International Journal of Interactive Multimedia and Artificial Intelligence

December 2024, Vol. IX, Number 1
ISSN: 1989-1660

unir LA UNIVERSIDAD EN INTERNET

"Once we learn how to write a program that can learn, then we can let it write its own programs. This brings an enormous satisfaction, as it mirrors our own learning process and the joy of understanding."
Marvin Minsky (The Society of Mind, 1986)

INTERNATIONAL JOURNAL OF INTERACTIVE MULTIMEDIA AND ARTIFICIAL INTELLIGENCE

ISSN: 1989-1660 –VOL. 9, NUMBER 1

EDITORIAL TEAM

Editorial: A New Chapter for IJIMAI

With great emotion and a profound sense of responsibility, I reach out to you, esteemed colleagues, readers, and contributors of the International Journal of Interactive Multimedia and Artificial Intelligence (IJIMAI), in this special issue that marks the end of a chapter for me as Editor-in-Chief. After eight volumes and more than fifty issues, I formally announce my departure from this role, passing the leadership to an accomplished and dedicated academic, Dr. Elena Verdú, who will assume the position of Editor-in-Chief.

The history of IJIMAI is a story of commitment, inclusion, and accessibility within the academic world. Since its founding, the journal has aimed to provide an open, free, and high-quality space for researchers and professionals in the fields of artificial intelligence and interactive technologies. The decision to offer open access to all was a clear commitment to democratizing knowledge and making scientific advances available to the entire community without financial barriers or access restrictions. Furthermore, IJIMAI has been developed, to date, on an open-source platform, underscoring our commitment to the values of transparency, accessibility, and global collaboration.

Throughout these volumes, IJIMAI has maintained a rigorous approach and an interdisciplinary vision, allowing us to address relevant topics from artificial intelligence and machine learning to interactive multimedia and complex systems. Each issue has been the product of a shared effort among authors, reviewers, and the editorial team, all working tirelessly to provide high-quality research and to promote scientific advancement in key areas.

These accomplishments would not have been possible without the collaboration of many individuals over the years. To the authors, whose dedication and hard work are the fundamental pillars of this publication, I extend my sincerest gratitude for trusting IJIMAI as a platform to disseminate your findings. Each article has contributed to the growth and development of our field.

To our editorial team and reviewers, who have upheld the highest standards of quality and rigor, I extend my deepest thanks. Your review and selection work has been essential in maintaining IJIMAI's scientific integrity and securing the recognition the journal enjoys today.

I would also like to express my heartfelt appreciation to my co-founders, Jesús Soto and Oscar Sanjuán, for their partnership and shared vision, without which IJIMAI would not be what it is today. From the very beginning, Jesús and Oscar demonstrated an unwavering commitment to scientific excellence and open access values, and their contributions have been essential at every step of our journey.

Furthermore, I would like to acknowledge and thank the unconditional support from the Universidad Internacional de La Rioja (UNIR), whose trust in this project and continued backing have been crucial to the journal's development. UNIR has been an invaluable supporter, enabling us to establish ourselves as a leading publication committed to innovation and knowledge advancement in artificial intelligence and interactive technologies.

It is with great satisfaction that I also announce the addition of Dr. David Camacho, Professor at the Universidad Politécnica de Madrid, as Advisory Editor for IJIMAI. Dr. Camacho's experience and academic networks will be an invaluable asset in further strengthening our visibility and international collaborations. I am confident that his contribution will play a crucial role in expanding the journal's reach and enriching our connections with researchers and scientific communities worldwide.

Finally, I am delighted to welcome Dr. Elena Verdú as the new Editor-in-Chief. Dr. Verdú brings a remarkable academic trajectory and an innovative vision that will undoubtedly lead IJIMAI to new heights. Under her leadership, I am confident that the journal will continue to establish itself as a beacon of academic excellence and accessibility in artificial intelligence and interactive multimedia.

Today, I bid farewell with gratitude and optimism, knowing that IJIMAI will continue to be an open and free platform, built on principles of transparency and a commitment to knowledge. I am certain that this new chapter, under Dr. Verdú's leadership and with Dr. Camacho's support, will be filled with accomplishments and advancements for our community.

Thank you all for your support and for joining us on this journey.

Sincerely,

Rubén González Crespo
Founder and Editor-in-Chief (Outgoing)
International Journal of Interactive Multimedia and Artificial Intelligence

DOI: 10.9781/ijimai.2024.11.004

TABLE OF CONTENTS

OPEN ACCESS JOURNAL

ISSN: 1989-1660

The International Journal of Interactive Multimedia and Artificial Intelligence is covered in Clarivate Analytics services and products. Specifically, this publication is indexed and abstracted in: *Science Citation Index Expanded, Journal Citation Reports/ Science Edition, Current Contents®/Engineering Computing and Technology.*

A Review of Bias and Fairness in Artificial Intelligence

Rubén González-Sendino[1], Emilio Serrano[1], Javier Bajo[1], Paulo Novais[2] *

[1] Ontology Engineering Group, Departamento de Inteligencia Artificial, ETSI Informáticos, Universidad Politécnica de Madrid, 28660 Madrid (Spain)
[2] ALGORITMI Research Centre/LASI, University of Minho, Braga (Portugal)

* Corresponding author: ruben.gonzalez.sendino@alumnos.upm.es (R. González-Sendino), emilio.serrano@upm.es (E. Serrano), jbajo@fi.upm.es (J. Bajo), pjon@di.uminho.pt (P. Novais).

Received 16 September 2022 | Accepted 29 September 2023 | Published 10 November 2023

Abstract

Automating decision systems has led to hidden biases in the use of artificial intelligence (AI). Consequently, explaining these decisions and identifying responsibilities has become a challenge. As a result, a new field of research on algorithmic fairness has emerged. In this area, detecting biases and mitigating them is essential to ensure fair and discrimination-free decisions. This paper contributes with: (1) a categorization of biases and how these are associated with different phases of an AI model's development (including the data-generation phase); (2) a revision of fairness metrics to audit the data and AI models trained with them (considering agnostic models when focusing on fairness); and, (3) a novel taxonomy of the procedures to mitigate biases in the different phases of an AI model's development (pre-processing, training, and post-processing) with the addition of transversal actions that help to produce fairer models.

Keywords

Bias, Fairness, Responsible Artificial Intelligence.

DOI: 10.9781/ijimai.2023.11.001

I. Introduction

The evolution of artificial intelligence (AI) has allowed humans to be heavily supported in the decision-making process of some application domains [1]. The high degree of independence that AI is capable of exhibiting can be problematic [2], [3], especially when humans are not in the loop [4]–[6]. Automatization of decisions can come at the cost of amplifying bias and creating feedback loops [7], [8]. One of the main reasons AI can produce unfair results is due to the data with which it has been trained [9].

Although the concept of *bias* is broad, this paper adheres to the following definition: "the systematic tendency in a model to favor one demographic group/individual over another, which can be mitigated but may well lead to unfairness" [9], [10]. Therefore, the next definition needed to understand the problem this paper studies is *Fairness*, which is defined as: " the absence of prejudice or favoritism towards an individual or a group based on its inherent or acquired characteristics" [9].

In the AI scope, incorrect predictions do not necessarily indicate that the model is unfair if its development was correct [11]. An unfair model is one whose decisions are biased toward a particular group of people. Moreover, biases cannot always be avoided. Thus, techniques must be used to mitigate their consequences, which aim to increase equality in the results. Data and models can be audited with *fairness metrics*, which are used to measure fairness between two groups or similar individuals. Furthermore, the categorization of methods for bias and unfairness mitigation depends on the phase of the AI model's development in which they are used. These phases are typically pre-training, training, and post-training.

This paper contributes with a systematic review of bias and fairness in artificial intelligence. The purpose of a systematic review is to provide a comprehensive summary of the literature available which is relevant to several research questions. The three questions addressed in this paper are: (1) What bias affects fairness?; (2) What are the metrics to measure fairness?; and, (3) How are biases mitigated? Beyond this systematic review and the taxonomy mentioned, the final goal of this paper is to help developers and researchers identify new biases and create fairer AI models.

This systematic review differs from others by addressing the three essential questions together. Previous systematic reviews have focused primarily on measurement and mitigation, complete systematic reviews on these fields are [12]–[14]. However, this review expands upon these works by including tools for auditing algorithms (Section VI) and guidelines for fair governance (Section VII). The bias affecting AI learning has been discussed in various papers, enumerating the different types or relating with recommender or scoring algorithms [9], [15], [16]. In this review, the bias has been organized into four general categories that apply to any type of AI system.

The content of the paper is organized as follows. Section II gives the background and motivation of the paper. Section III details the search criteria for the systematic review. Section IV shows the works retrieved for review. Then sections V, VI and VII respectively offer answers to the

Please cite this article as:
R. González-Sendino, E. Serrano, J. Bajo, P. Novais. A Review of Bias and Fairness in Artificial Intelligence, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 1, pp. 5-17, 2024, http://dx.doi.org/10.9781/ijimai.2023.11.001

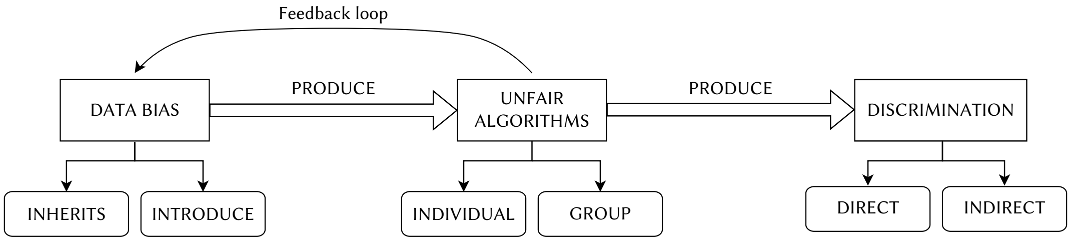


Fig. 1. Transformation of the bias in discriminatory results.

three research questions considered. In the context of the third research question ("How are biases mitigated?"), a taxonomy of the procedures to mitigate biases in the different phases of an AI model's development is presented. To finalize, Section VIII discusses the results and Section IX concludes and gives an overview of future work.

II. Background

The use of ML-based decision-making algorithms in organizations is increasing rapidly [17]. These algorithms may generate results that reflect, reproduce, and amplify structural inequalities. The results can be the product of unjust goals rooted in racist, sexist, hetero-normative, nationalist, or colonialist priorities [18].

These algorithms are becoming increasingly complex and *deep neural networks* (DNN) play an indispensable role in most AI-assisted tasks. These systems are "black boxes" due to the lack of transparency and explainability they exhibit. Therefore, DNNs can hide potential biases and present unexpected vulnerabilities [19], [20].

Due to previous problems, *Explainable Artificial Intelligence* (XAI) is becoming more necessary every day. Among others, XAI covers understandability, comprehensibility, interpretability, explainability, and transparency. These dimensions of XAI are essential to support and understand when discrimination occurs (explainability [21]–[24], interpretability [6], [25], [26] and transparency [1], [11], [20], [27]). XAI topics encourage *responsible AI* that considers fairness, privacy, accountability, ethics, transparency, security, and safety) [28].

Governments are focusing on explaining the decisions of autonomous systems to users. In Europe, new regulations give European citizens the right to have basic knowledge of the inner workings of automated decision-making models and to question their results [12], [29]. In Spain, there exists a similar law (BOE-A-2021-7840). This rule encourages the auditing of decisions made by an automated system. Since 2021, companies have had to inform their employers about the parameters, rules, and instructions that influence the algorithm's decisions.

There have been different incidents in which algorithms have produced unfair results. For example, the United Kingdom used an algorithm to infer the Advanced Level exam results for those students who could not take their tests[1]. In doing so, the algorithm considered the background of the students, their partners, or the school they attended. This resulted in disadvantaged ethnic minorities and people from poorer or disadvantaged backgrounds.

In the past, humans were discriminated against due to stereotypes, prejudice, or unintentional bias [30]. However, algorithms do not discriminate because they do not have the mental capacity to do so. In many cases, the problem is that human biases are transferred to the model by training data. The bias in a dataset can easily lead to an erroneous or discriminatory conclusion [31]. However, biases in AI algorithms can result not only because of issues with the training data but also from how algorithms learn over time and are used in practice [32].

Fig. 1 shows how biases are transformed into discrimination. Biases can be inherited (and later perpetuated) or introduced (and then exacerbated). The inherited bias perpetuates the existing inequality in the data structure [33]. Furthermore, bias can be introduced by assumptions in model implementation that exacerbates discrimination [34], [35]. The discrimination produced can be direct (disparate treatment) or indirect (disparate impact) [36]. Direct discrimination occurs when individuals receive less favorable treatment based on protected attributes such as sex, religion, or nationality [8]. Indirect discrimination occurs when people receive treatment based on inadequate factors. These factors are generally related to protected attributes [8]. Disparate impact is defined as a neutral rule that applies to everyone. However, the effect is more harmful to some people than to others [7].

Note that bias is subjective and related to the task. The healthcare results are not discriminatory if the diagnosis is based on sex-specific symptoms. However, the results of a hiring process could be discriminatory if it is sex-biased [7]. Note also that the problem of unfairness does not have to be addressed by necessarily reducing the use of AI. Sometimes, AI has been perceived as fairer than a human expert in the context of health and justice decisions [37]. AI decisions are made based on knowledge, unlike human decisions, which can be based on feelings.

To address bias, unfairness, and discrimination; AI must be audited following a procedure that involves: (1) the identification of potential biases that can affect fairness; (2) the selection of metrics to measure how fair AI is being; (3) and, the mitigation of the impact produced by these biases [32], [38], [39]. This paper conducts a systematic review of the state of the art with respect to these three key aspects.

III. Search Criteria for the Systematic Review

The main objective of this systematic review is to understand and analyze the fairness and bias in AI algorithms. To obtain a more detailed and comprehensive view of the field, the review examines the following three research questions (RQs):

- RQ1. What bias affects fairness?
- RQ2. What are the metrics to measure fairness?
- RQ3. How can biases be mitigated?

To strengthen the validity of the review, inclusion criteria (IC) for the studies included in this systematic review have been applied. These criteria and their justification are presented below.

- IC1. Studies must be peer-reviewed articles published in conferences, journals, press, etc.
- IC2. Studies must be conducted primarily in English.
- IC3. Studies must have been published since 2010.
- IC4. The abstract, introduction, and conclusion provide enough information.
- IC5. Studies that address the concept of fairness in artificial intelligence algorithms.
- IC6. Studies that provide enough information about bias as a product of unfairness.

[1] https://bit.ly/3dbxdbu

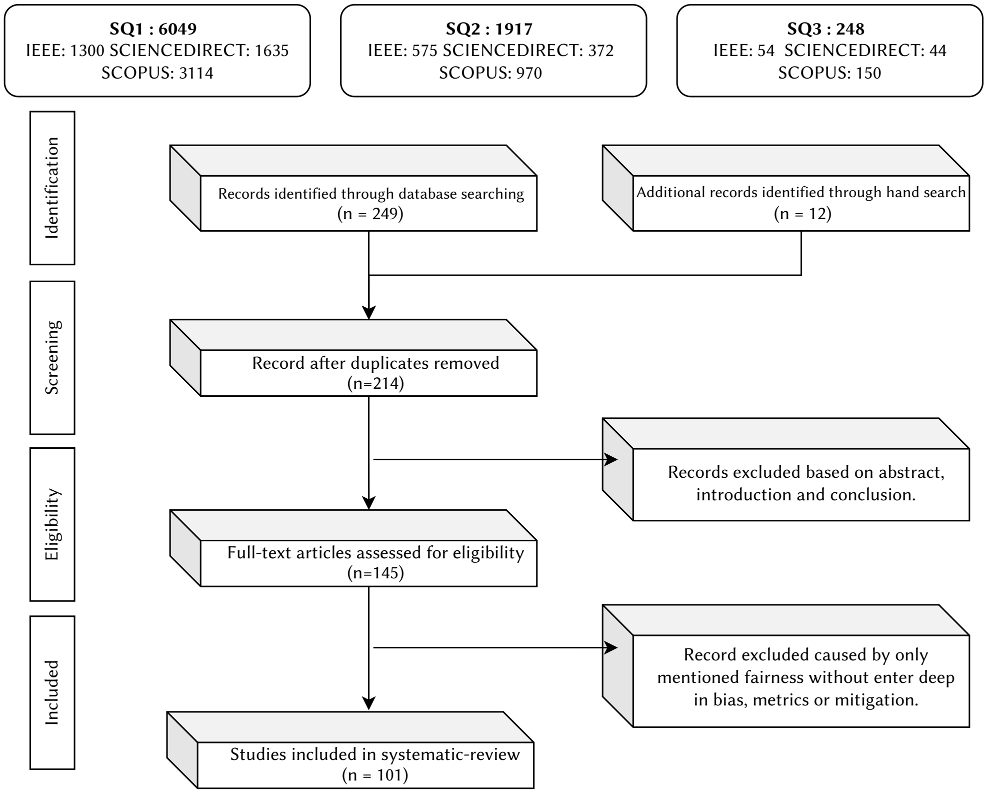

SQ1 : 6049
IEEE: 1300 SCIENCEDIRECT: 1635
SCOPUS: 3114
SQ2 : 1917
IEEE: 575 SCIENCEDIRECT: 372
SCOPUS: 970
SQ3 : 248
IEEE: 54 SCIENCEDIRECT: 44
SCOPUS: 150
Identification
Screening
Eligibility
Included
Records identified through database searching (n = 249)
Additional records identified through hand search (n = 12)
Record after duplicates removed (n=214)
Records excluded based on abstract, introduction and conclusion.
Full-text articles assessed for eligibility (n=145)
Record excluded caused by only mentioned fairness without enter deep in bias, metrics or mitigation.
Studies included in systematic-review (n = 101)

Fig. 2. PRISMA flow diagram depicting the flow of information through the different phases of the systematic review.

This systematic review seeks research on artificial intelligence that studies the unfair results produced by bias. The search queries are considered as a base including the terms: fairness, artificial intelligence, machine learning, and bias. Furthermore, the terms FAT (fairness, accountability, and transparency) and FATE (fairness, accountability, transparency, and explainability) are considered relevant to find related research. XAI and responsible AI are excluded because these topics are considered transversal to the research questions studied and therefore can include noise in the systematic review.

The systematic review focuses on the latest biases, metrics, and mitigation techniques. For this reason, the search query applies a filter by date; only articles published since 2010 will be considered. The scientific libraries used for the collection of articles will be Science Direct, Scopus, and IEEE.

The search query (SQ) will be applied only to: the title, abstract or author-specified keywords.

- SQ1 is ((‘artificial intelligence’ or ‘machine learning’) and (‘fair’ or ‘fat’ or ‘fate’ or ‘bias’ or ‘fairness’ or ‘unfair’)).
- SQ2 is ((‘artificial intelligence’ or ‘machine learning’) and (‘fair’ or ‘fat’ or ‘fate’ or ‘fairness’ or ‘unfair’)).
- SQ3 is (‘artificial intelligence’ or ‘machine learning’) and (‘fair’ or ‘fat’ or ‘fate’ or ‘fairness’ or ‘unfair’) and (‘bias’).

The three queries are used to filter from a complete set to a subset that collects the desired papers $SQ1 \supset SQ2 \supset SQ3$. The articles filtered by *SQ3* contain both topics: fairness and bias. This requirement reduces the noise caused by the broadness of the concept of bias.

IV. Retrieved Works for the Systematic Review

The three search queries used to find related research produce distinct results in terms of quantity. Fig. 2 shows the number of results for each search.

SQ1 returned thousands of studies, most of them related to an unbalanced dataset without linking it with fairness. A large number of the SQ2 results talked about fairness in algorithms, as a top-level definition without diving deep into the topic. Finally, SQ3 is the search query used to filter and read information related to fairness and bias for machine learning algorithms.

Note that the solution to bias could be similar to the solution to having unbalanced data. However, in the latter scenario, the purpose is to improve the accuracy of the model while in the former the objective is to reduce unfairness. Several studies tried to find a trade-off between accuracy and fairness.

Fig. 2 shows a flow chart with the filters applied to reduce the number of papers included in this systematic review. After applying the search queries and the inclusion criteria, the number of research works considered is 101.

Fig. 3 shows that the number of publications related to this research is growing significantly and has gained relevance in the last five years. For this reason, having skipped papers before 2010 does not seem to be a problem.

Research can be grouped into three main categories: fields in which there is a concern with bias; algorithms in which the fairness metrics are trying to be mitigated; and theoretical studies about bias and its mitigation. The fields help to understand where researchers face unfairness, thus discovering why biases affect the fields. Additionally, understanding the deficiencies of the learning algorithm could help to understand the complexity of obtaining a fair output.

Table I shows the fields and algorithms most popular in this review. Health is the most relevant field in which research seeks equity. The next most crucial field is recruitment and education. This focus of the studies is understandable because of the importance of decisions in these fields. Table I also details the main algorithms used in the retrieved works in the specialized literature. Neural networks are the most

TABLE I. Important Field and Algorithms Where Some Studies Are Centered

Field	Nº
Health [10], [19], [32], [40]–[46]	10
Education [9], [20], [47], [48]	4
Recruitment [17], [49], [50]	3
Travel [33]	1
Manufacturing [2]	1
Laws [1]	1
Public service [51]	1
Rent house [38]	1
Deep Neural Networks [33],[34], [52]–[55]	5
Ambient intelligence [56]–[60]	5
NLP [7], [61]–[63]	4
Computer Vision [31], [64],[65]	3
GAN [66], [67]	2
Decentralizing Learning [68], [69]	2
Support Vector Machine [70], [71]	2
Decision tree [20], [72]	2
Adaptive models [32]	1
XGBoost [22]	1

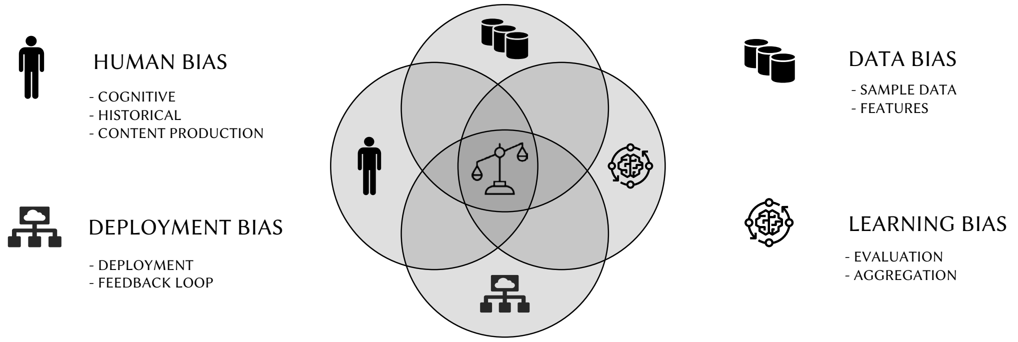


Fig. 4. An overview of bias impacting fairness.

widespread learning paradigm. The use of these networks is generally: decision-making, recommendation, scoring, or classification.

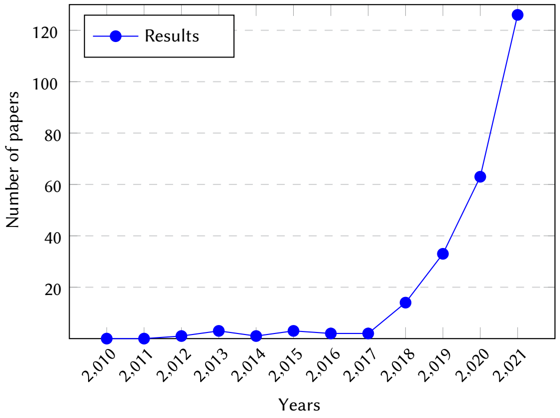


Fig. 3. Graph showing the evolution of publications related to bias and unfairness in artificial intelligence algorithms.

There are publications that address the topics of fairness and bias from a theoretical and transversal perspective in artificial intelligence. The majority of these papers talked about mitigation techniques [12]–[14], [22], [23],[73]–[76], additionally the topics of fairness metrics [22],[23], [29], [74], [76] and types of bias [15], [16], [73],[77], [78] are covered.

V. RQ1: What Bias Affects Fairness?

This section provides an overview of the biases that affect fairness. Fig. 4 shows the four phases in which the biases can be grouped. The groups to identify biases are linked to the model life cycle: production (human) bias, data bias, learning bias, and deployment bias.

The relevant consequence of biases is discrimination, reflected in the tendency to favor one individual or group over another [10]. Fig. 1 illustrates the transformation of bias into discrimination using an unfair model.

A. Human Bias

Human beings are the main factor that produces bias. For this reason, the production bias group in Fig. 4 is called *human bias*. Human-made decisions that are reflected in the data or in the model. Therefore, eliminating bias in machine learning and Artificial Intelligence without addressing the pressing concerns about bias in humans is not possible [76].

The human bias group in Fig. 4 collects how biased data are generated. Human decisions can lead to unfairness in a number of steps in AI development discussed below: data management, learning, and model deployment. These steps will be compiled in the other groups. Human biases are divided into two main subgroups: cognitive bias, and behavioral bias. Note that the information is susceptible to variation over time, producing *temporal bias* [77].

- *Cognitive bias* is a deeply ingrained part of human decision making [73], which transfers prejudices to labels [79]. Machine learning algorithms use human judgments as training data, so they propagate these biases. This *historical bias* arises even when the data is perfectly measured and sampled, for example, reinforcing a stereotype [7].
- *Behavior bias* produces distortions from reality or other applications according to user connections, activities, or interactions [9]. Furthermore, unconscious bias could be produced by *content creation*, because the way a child or an adult expresses themselves is different, in the same way as if you compare by sex or race [61]. Content creation bias is also produced when users are guided by norms or functionalities [9], and sometimes these interactions are led by an AI system [32]. Bias production in the future will be affected by unfair systems, generating new data to be used in future learning [9].

B. Data Bias

Data bias focuses on the factors that induce a biased dataset. The main tasks where data bias may occur are acquisition, querying, filtering, transforming, and cleaning [9].

Data bias is generally a distortion in *sampled data* that compromises its representatives [15], [41]. In other words, sample bias is produced when the train and test data do not represent or under-represent a population segment. Therefore, modelers play a critical role in producing data bias by including or discarding data [35], and even labeling the data [9].

The *representation* is not the only problem related to the data. *Features* could reflect discrimination by sensitive attributes: race, sex, age, socio-economic data, education, neighborhood, etc. These features are generally not legitimate for decision-making [80]. Typical sensitive variables are those collected by the General Data Protection Regulation (GDPR).

Fairness is not only attributable to protected attributes. Proxy attributes can be exploited to derive sensitive features [14]. Generally, protected characteristics and targets are highly correlated, resulting in good accuracy at the cost of diminishing fairness metrics [33]. The correlation applied as causality could lead to bias [9].

C. Learning Bias

Learning biases are produced in model training. This step is compounded by the difficulties of understanding and explaining the results. Typically, models learn a correct statistical pattern in favor of the majority over minorities [15], [33], leading to *aggregation* bias that amplifies the disparities between different examples in data samples [7]. At this point, the model that learns from generated experience, such as reinforcement learning, could become biased over time [32].

The performance of the model should be evaluated after each training loop. *Evaluation bias* occurs when the selected metrics are not appropriate, for example, the use of general vs. subgroup accuracy [77], and when the representation in the test data does not reflect reality.

D. Deployment Bias

Deployment bias may occur in the deployment and use of the model. The algorithm makes decisions based on patterns learned from the data. Therefore, *the deployment* of a model in a different scenario with respect to data could lead to unfair results [10]. Feedback loop is the result of the introduction of a new discriminatory decision in the data [16].

VI. RQ2: What Are the Metrics to Measure Fairness?

This section collects approaches to measure fairness. Measurement of fairness gives a quantification of the problem for further mitigation. Although these issues are most apparent in the social sciences, where fairness is interpreted in terms of the distribution of resources across protected groups, the management of bias in source data affects a variety of fields. Any domain involving sparse or sampled data is exposed to potential bias [26].

Typically, metrics used to measure fairness are divided into two groups. On the one hand, metrics measure and find the difference in equality between two selected groups. On the other hand, metrics can compare results between similar individuals whose results are disparate [81]. The final goal of both is to find discriminatory inputs [81].

- *Group Fairness or Disparate Impact.* Each group identified in the dataset receives an equal fraction of a possible outcome (applies to both positive and negative outcomes) [29], [82]. In other words, different sensitive groups should be treated equally. The two groups usually are called the Unprivileged Group (UG) and the Privileged Group (PG).
- *Individual Fairness or Disparate Treatment.* Individuals who belong to different sensitive groups with similar characteristics should be treated similarly [29], [82]. For example, applicants with the same qualifications during job applications should not be discriminated against based on their sex or race. Some positions highlight that individual fairness cannot be a definition of fairness due to: insufficiency of similar treatment, systematic bias and arbitrators, and prior moral judgments [83].

The most common metrics used to measure fairness are shown in Table II. Those metrics require ground-truth data. Other metrics are used for unsupervised problems such as Fairness Demographic Parity, Point-wise Mutual Information, Kendall Rank Correlation, t-test, and Log-likelihood Ratio [84].

A less extended classification divides algorithms into: statistics based on predicted outcomes, statistics based on predicted and actual outcomes, statistics based on predicted probabilities and actual outcomes, and similarity-based and causal reasoning [7], [36], [85]. The two most employed are: statistics based on predicted outcomes and statistics based on predicted and actual outcomes.

Statistics based on predicted outcomes can be defined as statistical parity, conditional statistical parity, and predictive equality. Furthermore, statistics based on predicted and actual outcomes can be described as calibration within groups, balance for the negative class, and balance for the positive class. Not all algorithms can satisfy all conditions simultaneously. The objective is a trade-off between the ability to classify accurately and the fairness of the resulting data [36].

A significant thing about the status of bias detection is that current strategies for detecting biases are often customized for a problem, dataset, or method [88]. This affects their generalization [88]. (1) For group fairness, there exist simple studies that use Receiver Operator Characteristics (ROC) curves for each demographic group [10], [31]. Other traditional metrics used to calculate fairness are standard deviation and skewed error ratio [65]. (2) For individual fairness, there exists Procedural Fairness [29] and Consistency Metric [22]. Procedural fairness ensures that the algorithm does not use sensitive features for prediction. Consistency Metric compares the prediction of a certain individual with the predictions of its k-nearest neighbors. The bias disparity is a concept introduced in Aequitas [89]. This bias is calculated by comparing the metric for a given group with the metric of the reference group. The metrics that could be calculated are: predicted positive, total predictive positive, predicted negative, predicted prevalence, false positive, false negative, true positive, false negative, false discovery rate, false omission rate, false positive rate, and false negative rate.

Finally, the existence of agnostic models helps to comply with responsible artificial intelligence. The most common agnostic models include explainability methods. There exist specific agnostic models that focus on helping to improve fairness in the different development phases: AIF360 [23], FairLearn [44], [90], LFIT [50], Aequitas [89], LimeOut [29], MAML [67], the What-If toolkit (WIT) [91], and Audit AI [92].

The previous toolkits help to audit machine learning models for discrimination and bias. The most popular are Aequitas (Carnegie Mellon University), AIF360 (IBM), FairLearn (Microsoft) and WIT (Google). Some of these tools, in addition to measure, also help with mitigation. This is the case of FairLearn and AIF360.

Aequitas and WIT are especially suitable as audit tools. WIT provides a graphical interface in which the behavior of an algorithm can be tested visually. The tool integrates the fairness indicator developed in TensorFlow. Aequitas runs a full report on biases. This report is expected to be used by developers, analysts, and policymakers.

VII. RQ3: How to Mitigate Bias?

As a result of the systematic review of this research question, a taxonomy has been proposed to aggregate bias mitigation procedures. Fig. 5 displays this new taxonomy which is considered from the point of view of data science. The stages where mitigation techniques can be applied include pre-training, training, and post-training.

- Mitigating bias in the pre-training phase is the most effective manner of correcting bias since it transforms the dataset. However, bias may appear after training, hindering developers from dealing with it in the first iteration of the process [73].
- Training is the most efficient stage for handling bias. These methods are often unsupervised and do not involve adulterating the underlying data set [73]. Not including sensitive features such as gender or race is not enough to mitigate discrimination, considering that other derivative features are introduced. Instead, adding fairness to the objective function is more efficient [93].
- Post-training is an ideal phase to calculate most of the previously revised metrics [73]. However, mitigating biases in this phase should be the last option [67].

There is a limitation with pre-processing and post-processing because manipulating the data leads to an outcome that may not be realistic due to the perturbation of the original distribution [53].

A fair governance category is also included in the taxonomy, where mitigation is possible without applying complex algorithms.

A. Pre-Training

Pre-processing modifications could be made to the sample or features and labels to produce fair data that neutralize discriminatory effects [65]. However, this approach cannot eliminate discrimination that may come from the algorithm itself [7]. The taxonomy shown in Fig. 5 reflects the three main techniques studied in the literature: resampling, fair representation, and re-weighting.

1. Resampling

Resampling is used to change the size of the data set that affects the distribution without transforming the data. Resampling methods are divided into undersampling and oversampling [65], [67]. Undersampling techniques are based on eliminated samples from the dataset; meanwhile, oversampling means generating (or repeating) data samples to augment the original dataset.

These techniques have been transferred from data-balancing problems to the fairness domain. In fairness mitigation, different algorithms are tested for data augmentation, while techniques for undersampling are less popular [94]. Successive data augmentations may be computationally expensive if the dataset contains many features [79].

The two hegemonic approaches to oversampling are: the Synthetic Minority Oversampling Technique (SMOTE); and, the Generative Adversarial Networks (GANs) [65]. GANs have been used to produce synthetic tabular data to improve demographic parity [66], allowing fairness to be increased while maintaining precision in prediction.

In addition to altering the number of samples, another approach to improve fairness can be to reduce the number of features in the data by *feature selection.* A simple method is to remove sensitive features that could produce bias in prediction. However, this is not enough, since protected attributes could be encoded or correlated with other features [12].

2. Fair Representation

Fair representation is obtained by eliminating information that can link a person to a protected group [73]. *Learning fair representation* (LFR) is a popular algorithm for finding a latent representation that encoded data while preserving fairness [22]. Protected information can be hidden or explicit, giving more or less weight to its representation [76]. However, LFR improves fairness at the cost of complicating the explainability of the results [74].

3. Re-Weighting

Re-weighting is the method more widely used to transform the data by modifying the weight in the data set [75]. Note that not all learning algorithms accept weighted samples [73]. Re-weighting means that certain instances from a privileged group, more likely to have a favorable outcome, will get a lower weight. Similarly, instances of an unprivileged group will receive a higher weight [22].

4. Other Categories of Bias Mitigation in Preprocessing

In addition to the three main categories explored above, less popular methods used to mitigate bias in the preprocessing phase include: Privileged Group Selection Bias (PGSB) [13], [14], disparate impact remover [23],[76], and optimized preprocessing [23], [74].

B. Training

When mitigating biases in training time, algorithms are modified to improve fairness rather than just precision. The advantage of addressing biases in this phase see Fig. 5, is that data and prediction can be used to evaluate fairness. Regularization and adversarial training are the most common methods for this purpose according to the revised literature. Other emerging approaches are: decentralized learning, fair linear regression, fair-n, DeepFair, multimodal models, and fairlet clustering. These approaches are discussed below.

1. Regularization

Regularization is a well-known technique in machine learning. Regularization is used to correct underfitting or overfitting when training the model. This method can also be used to mitigate biases and unfairness [73]. In contrast, adding regularization methods to a machine learning model can complicate the explanation and interpretation of its results [22]. Regularization for mitigating biases can be: implicitly adding constraints that disentangle the association between model predictions and sensitive attributes; or explicitly adding constraints by updating the model loss function to minimize the performance difference between different protected groups [65].

Regularization methods in the loss function of deep neural networks can help reduce the difference in prediction disparity between different groups [33]. Regularization can also penalize high correlations between sensitive attributes and outcomes. The following reports [13], [14], [95] employ L2 regularization to weight examples equally in several machine learning models, such as support vector machines (SVMs) and logistic regressions (LRs).

2. Adversarial Training

In the field of AI, *Adversarial Learning* is a technique in which multiple neural networks compete with each other to improve the predictive accuracy [96]. The fairness of machine learning models can be improved by mitigating bias through the use of adversarial learning; this process is called *adversarial debiasing* [31].

Adversarial debiasing involves training two neural networks where one network learns to predict the outcome, and the other network identifies and removes any biases in the training data that could affect the prediction of the first network. The second network, also known as the "adversary", attempts to find and exploit weaknesses in the first predictions, thus forcing the first network to become more robust and resistant to bias [97].

TABLE II. The Most Popular Metrics to Measure Fairness

Metric name	**Target**	**Definition**
Equal Opportunity Difference (EOD)[22], [23], [29], [74],[76], [82], [86]	Group	Measures the difference in true positive rates (TPR) between an unprivileged group and a privileged group. $\mathrm{TPR} = \left[\frac{\text{True Positive (TP)}}{\text{TP + False Negative (FN)}}\right]$ (1) $\mathrm{EOD} = \mathrm{TPR}_{\mathrm{UG}} - \mathrm{TPR}_{\mathrm{PG}}$ (2)
Odds Difference (OD) [22], [23], [74], [76], [79]	Group	Computes the difference of false positive rate (FPR) and true positive rate (TPR) between unprivileged and privileged groups. $\mathrm{FPR} = \left[\frac{\text{False Positive (FP)}}{\text{FP + True Negative (TN)}}\right]$ (3) $\mathrm{OD} = (\mathrm{FPR}_{\mathrm{UG}} - \mathrm{FPR}_{\mathrm{PG}}) + (\mathrm{TPR}_{\mathrm{UG}} - \mathrm{TPR}_{\mathrm{PG}})$ (4)
Statistical Parity Difference (SPD) [22], [23], [74], [75],[75], [76], [79], [86]	Group	Calculates the difference in the probability of favorable results (Predicted as Positive (PPP)) between the unprivileged group and the privileged group. $\mathrm{PPP} = \left[\frac{\text{TP + FP}}{\text{Total Population (N)}}\right]$ (5) $\mathrm{SPD} = \mathrm{PPP}_{\mathrm{UG}} - \mathrm{PPP}_{\mathrm{PG}}$ (6)
Disparate Impact (DI) [22], [23], [74],[76]	Group	Compares the proportion of individuals who receive a positive output for two groups: an unprivileged group and a privileged group. $\mathrm{DI} = \frac{\mathrm{PPP}_{\mathrm{UG}}}{\mathrm{PPP}_{\mathrm{PG}}}$ (7)
Theil Index (TI) [22],[23], [87]	Group/ Individual	Subclass of the generalized entropy index (using alpha = 1). The entropy index is a measure of inequality in a group or individual with respect to the fairness of the algorithm outcome. $\mathrm{TI} = \frac{1}{N}\sum_{i=1}^{N}\frac{b_i}{\mu}\ln\frac{b_i}{\mu}, b = predicted - labeled$ (8)

In adversarial debiasing, the goal is to reduce evidence of any biases related to protected attributes in the predictions [73]. Evidence of protected attributes can be reduced, and prediction accuracy can also be improved in certain cases [75]. Scores between different demographic groups can be balanced, promoting: demographic parity, equality of odds, and equality of opportunity [97].

3. Emerging Approaches

Decentralizing the learning is an application of the blockchain strategy to machine learning models where models are built with a distributed and collaborative approach. In this scheme, some methods that have been examined from the point of view of fairness in the literature are: Swarm Learning (SL) [69] and Federated Learning (FL) [68].

The *Fair linear regression* and *Fair-n* approaches are useful for cases with more than one sensitive variable. *Fair linear regression* is based on the Hilbert-Schmidt independence criterion. This allows it to deal with several sensitive variables simultaneously [93]. *FaiR-N* introduces fairness and robustness regularization techniques to the loss function based on an approximation of the distance of data points to the decision boundary during training [53].

Fairlet clustering can be used in cases with unsupervised data [71], where detecting bias is a complex task. In clustering problems, fairness is defined in terms of consistency in that the balance ratio of data with different sensitive attribute values remains constant for each cluster.

DeepFair is a solution for a recommender system [55]. The recommender system relies on Collaborative Filtering (a set of the user's preferences on the items). This amount of data affects minorities negatively. The solution proposes a Deep Learning based on Collaborative Filtering that provides recommendations with an optimal balance between fairness and accuracy.

Multimodal models can understand and process information from multiple heterogeneous sources of information that can help reduce or correct bias and unfairness [49]. However, in multimodal model, detecting the origin of the bias is a very challenging task [17].

C. Post-Training

When addressing bias in this phase, the results of the model are modified by correcting decisions that could harm the fair representation of different subgroups in the final decision process [7]. As shown in Fig. 5, the most commonly used methods in post-training include: equalized odds, calibrated equalized odds, and reject option classification.

- *Equalized odds* adds a post-learning step to determinate optimal probabilities to change output labels. Equalized odds enforce fairness and precision [22], [23], [65],[73]–[75].
- *Calibrated equalized odds*, starting from the score outputs of a calibrated classifier, optimizes the probabilities with which to change the output with an equalized odds objective [23], [65], [73]–[75].
- *Reject option classification* gives favorable outcomes to protected unprivileged groups and unfavorable outcomes to privileged ones. This method uses a confidence band around the decision boundary with the highest uncertainty [22], [23], [65], [73]–[75].

D. Fairness Governance Practices

Fair governance minimizes biases while avoiding mitigation methods revised above for pre-training, training, and post-training. As shown in Fig. 4, the three main areas in which unfair results are reduced are: team, data, and models.

1. Team

This section refers to the people involved in developing an artificial intelligence algorithm. The impact of the modeler is not the only one. For example, imposed precision requirements may go against fairness indicators.

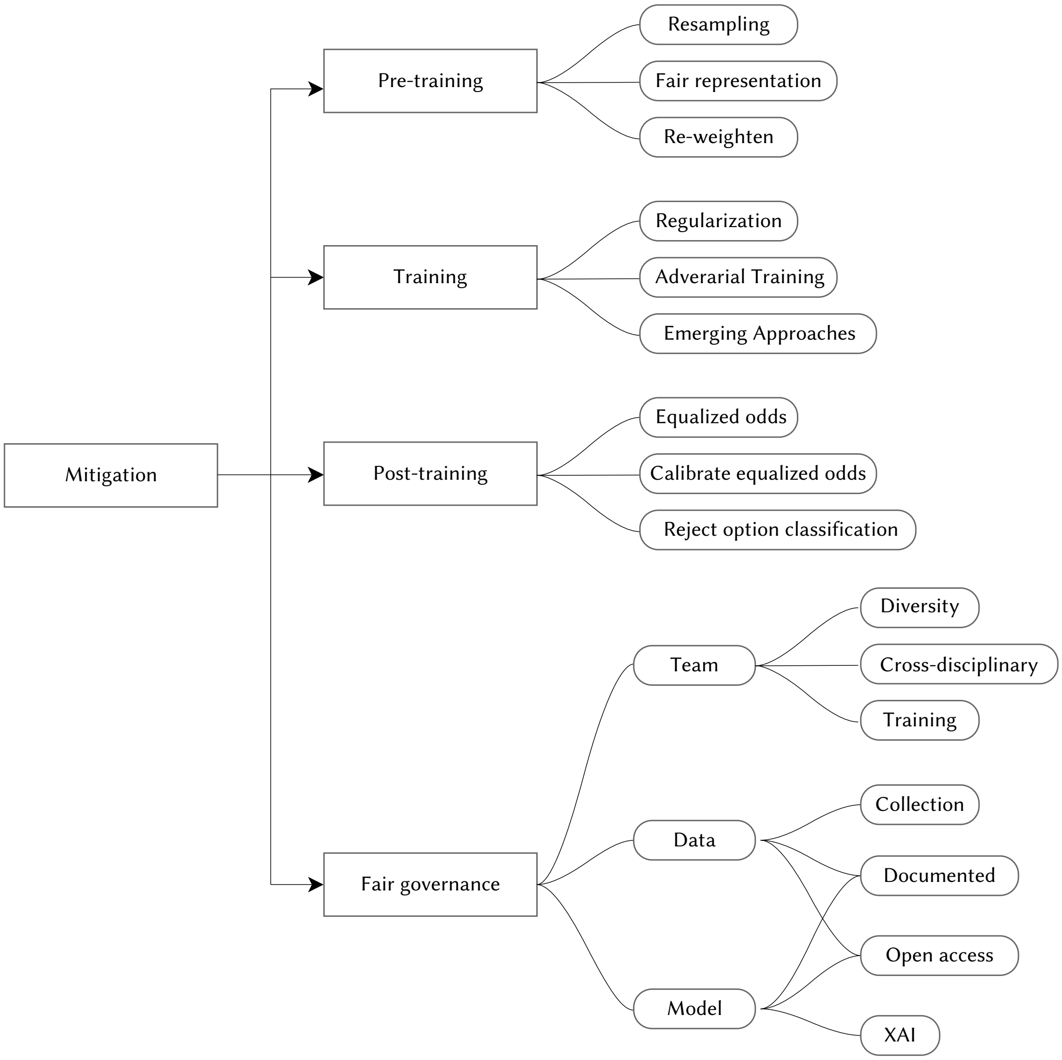


Fig. 5. Taxonomy of bias mitigation.

Diversity is the first dimension of the teams that needs to be improved [98]. Furthermore, this diversity has to affect all levels of the hierarchy [4]. Until then, algorithms and their associated biases will become mirrors of structural discrimination rather than bridges to opportunity, equality, and efficiency [99].

Creating diverse teams, as well as *cross-disciplinary* teams of data scientists and social scientists [5], is also essential to reducing bias and unfairness.

To further improve AI development, continuous training in fairness and ethics is recommended for team members, and stakeholders [4]. In addition, cultivating emotional intelligence to process better, identify, and confront discordance [6]. Tools, such as packages to calculate and audit bias, can be an important aid in addressing and standardizing this problem [9].

2. Data

The data is where the knowledge resides, and the algorithms will learn that information. Important keys at this point could be selecting features, transformations, or data labelling.

Data collection can be improved by prioritizing data on sex, race, ethnicity, etc. As a result, complete information related to sensitive features is available in the datasets. Therefore, knowledge of the sensitive issues present in real life is improved. This allows representative problems to be reduced [5].

Data-based decisions must be *documented* with original priorities and the necessary annotations [6]. This allows answering which data have been used consistently with inclusion patterns [100].

Opening access to train the data used to build the model leads to better transparency and trustworthiness. In addition, this allows third parties to detect possible biases [9].

3. Model

This topic gives suggestions for the training and evaluation phases. Furthermore, during the deployment it is essential to understand and explain the results. The model perpetuates or creates bias independently of the origin of this bias. Fig. 4 shows that under the fair governance category of bias mitigation methods, the “model” and “data” sections share the “documented” and “open access” recommendations explained above.

Models open to the community would allow their testing to detect new biases and demonstrate their efficacy [9]. Also, documentation of assumptions on parameters or metrics should be transparent and available [6].

Furthermore, applying *Explainable Artificial Intelligence* (XAI) allows understanding the model results and the importance of each feature in the predictions [4]–[6].

VIII. Discussion

This section explains the results obtained in this systematic review and the answers to the research questions revised. Section III has detailed the search criteria for this systematic review. These criteria pursue robust and unbiased results. Among others, only peer-reviewed research works have been considered, which is the first of the six inclusion criteria discussed. Although peer review is not a guarantee of a good level of confidence in the published results, it is seen as the foremost process for research validation [101].

A. RQ1: What Bias Affects Fairness?

Regarding RQ1, results were quite broad because the term *biases* encompasses different concepts within AI, including poorly balanced datasets which can lead to performance issues. However, this paper specifically focuses on biases that are related to injustice. Many studies on this question analyze biases for specific use cases, such as classification or recommendation. Other publications listed biases that could appear in the training cycle (from data acquisition to algorithm deployment).

As a result of this question, Fig. 4 groups biases in the following steps of an algorithm development process: Human Bias (Data Generation), Data Bias, Learning Bias, and Deployment Bias. In all the groups, the focus is placed on the human factor, which is responsible for the decisions made.

The critical step detected in this point is Human Bias. As discussed earlier, this is where historical data for training is generated in a way that can introduce biases. The bias in the following steps could be reduced or eliminated by reducing the cognitive bias presented in the dataset.

B. RQ2: What Are the Metrics to Measure Fairness?

The topic of metrics has only appeared in a few of the retrieved publications, being the most widespread metrics detailed in Table II. Another clear result is that the most used metrics are the ones that focus on groups. Moreover, this paper also reviews metrics that have recently emerged (generally customized for a specific case). Tools to audit algorithms have also been added to complete the answer to this research question.

As a result of this review, an interesting gap in the literature has been found. None of the revised metrics helps to detect variables or values that can cause unfairness. Therefore, it is necessary to know which groups are privileged and unprivileged to measure the differences between selected groups.

Applying these metrics to the entire dataset requires a high computational cost. Moreover, these metrics cannot be used with all variables because some of them, without being a discriminatory feature, can correctly segregate the samples.

C. RQ3: How to Mitigate Bias?

In the last question, the aim was to find techniques that would help mitigate biases. This is the question that most of the papers covered have focused on.

As a result of the review presented in this paper, mitigation techniques have been grouped into a taxonomy in Fig. 5. This taxonomy contains four main categories: Pre-training, Training, Post-training, and Fair Governance.

Most of the retrieved works focus on correcting data (pre-training) or improving learning (training). The pre-training algorithms include: Resampling, Fair Representation, and Re-weighting. The training algorithms consider: Regularization, Adversarial Training, and Emerging Approaches. Training algorithms are where the most varied solutions are developed. At this category, the main goal is to maintain accuracy while improving fairness. On the other hand, focusing on the output (post-training) is a less popular approach.

Some publications address their research to detecting transversal actions to mitigate unfairness. These actions attempt to improve the development ecosystem. The Fair Governance category contains actions that reduce bias when applied to: Teams (Diversity, Cross-disciplinary, and Training), Data (Collection, Documented, and Open Access), or Models (Documented, Open Access, and apply XAI techniques). This category can reduce bias in the Data Bias, Learning Bias, and Deployment Bias. Thus, the Fair Governance category is essential to have an impact on fairness.

IX. Conclusions and Future Work

This paper has established a systematic review to answer three research questions.

The first question focuses on understanding the origin of unfair results in AI models. As a result of the review, a comprehensive analysis of the type of biases that affect fairness was produced.

The second question explored equity metrics to detect discrimination in data or models. Studies show that quantifying this issue is very complex with the current state of the art. The review identifies two main factors needing improvement: obtaining generalized measures and automatically detecting sensitive features. The paper contributes to a compilation of the most popular and novel metrics to measure fairness.

The last question was aimed at obtaining information on how to mitigate the effects of bias in AI models. According to the extensive specialized literature reviewed, this mitigation is still a complex and imprecise task. More importantly, reducing bias can change learning and obtain undesirable results. Among others, the results could not represent reality when the algorithm avoids historical discrimination. A taxonomy that aggregates the different mitigation techniques depending on where they are applied is this paper's third and main contribution.

Future work should address the development of a fairness-by-design standard for developing AI models. In addition, the detection of feature bias should be automated at least for sensitive variables or their derivatives. Finally, an indicator of responsible AI development is needed beyond the use of performance metrics.

Acknowledgment

This work has been granted by the "EICACS (European Initiative for Collaborative Air Combat Standardisation)" project of the Horizon Europe programme of the European Commission, under grant agreement No. 101103669. The work of Paulo Novais is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia within project DSAIPA/AI/0099/2019.

References

[1] R. Kennedy, "The ethical implications of lawtech," in *Responsible AI and Analytics for an Ethical and Inclusive Digitized Society*, 2021, pp. 198–207, Springer International Publishing.

[2] S. Camaréna, "Engaging with artificial intelligence (ai) with a bottom-up approach for the purpose of sustainability: Victorian farmers market association, melbourne australia," *Sustainability*, vol. 13, no. 16, 2021, doi: 10.3390/su13169314.

[3] S. Strauß, "Deep automation bias: How to tackle a wicked problem of ai?," *Big Data and Cognitive Computing*, vol. 5, no. 2, 2021, doi: 10.3390/bdcc5020018.

[4] A. Nadeem, O. Marjanovic, B. Abedin, "Gender bias in ai: Implications for managerial practices," in *Responsible AI and Analytics for an Ethical and Inclusive Digitized Society*, 2021, pp. 259–270, Springer International Publishing.

[5] S. Parsheera, "A gendered perspective on artificial intelligence," in *2018 International Telecommunication Union Kaleidoscope: Machine Learning for a 5G Future, Santa Fe, Argentina, November 26-28, 2018*, 2018, pp. 1–

[6] 7, IEEE.

[7] T. K. Gilbert, Y. Mintz, "Epistemic therapy for bias in automated decision-making," in *Proceedings of the 2019 Conference on AI, Ethics, and Society*, New York, NY, USA, 2019, p. 61–67, Association for Computing Machinery.

[8] D. A. da Silva, H. D. B. Louro, G. S. Goncalves, J. C. Marques, L. A. V. Dias, A. M. da Cunha, P. M. Tasinaffo, "Could a conversational ai identify offensive language?," *Information*, vol. 12, no. 10, 2021, doi: 10.3390/info12100418.

[9] C. Zhao, C. Li, J. Li, F. Chen, "Fair meta-learning for few-shot classification," in *2020 IEEE International Conference on Knowledge Graph, Online, August 9-11, 2020*, 2020, pp. 275–282, IEEE.

[10] R. S. Baker, A. Hawn, "Algorithmic bias in education," *International Journal of Artificial Intelligence in Education*, pp. 1052–1092, 12 2021, doi: 10.1007/s40593-021-00285-9.

[11] R. R. Fletcher, A. Nakeshimana, O. Olubeko, "Addressing fairness, bias, and appropriate use of artificial intelligence and machine learning in global health," *Frontiers in Artificial Intelligence*, vol. 3, 4 2021, doi: 10.3389/frai.2020.561802.

[12] M. Loi, A. Ferrario, E. Viganò, "Transparency as design publicity: explaining and justifying inscrutable algorithms," *Ethics and Information Technology*, vol. 23, pp. 253–263, 9 2021, doi: 10.1007/s10676-020-09564-w.

[13] E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M. E. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis, I. Kompatsiaris, K. Kinder-Kurlanda, C. Wagner, F. Karimi, M. Fernandez, H. Alani, B. Berendt, T. Kruegel, C. Heinze, K. Broelemann, G. Kasneci, T. Tiropanis, S. Staab, "Bias in data-driven artificial intelligence systems—an introductory survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, 5 2020, doi: 10.1002/widm.1356.

[14] D. Pessach, E. Shmueli, "Improving fairness of artificial intelligence algorithms in privileged- group selection bias data settings," *Expert Systems with Applications*, vol. 185, 12 2021, doi: 10.1016/j.eswa.2021.115667.

[15] D. Pessach, E. Shmueli, "A review on fairness in machine learning," *Association for Computing Machinery: Computing Surveys*, vol. 55, feb 2022, doi: 10.1145/3494672.

[16] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, "A survey on bias and fairness in machine learning," *Association for Computing Machinery: Computing Surveys*, vol. 54, no. 6, pp. 115:1–115:35, 2021, doi: 10.1145/3457607.

[17] S. Khenissi, B. Mariem, O. Nasraoui, "Theoretical modeling of the iterative properties of user discovery in a collaborative filtering recommender system," in *Proceedings of the 14th Conference on Recommender Systems*, New York, NY, USA, 2020, p. 348–357, Association for Computing Machinery.

[18] A. Peña, I. Serna, A. Morales, J. Fiérrez, "Faircvtest demo: Understanding bias in multimodal learning with a testbed in fair automatic recruitment," in *International Conference on Multimodal Interaction, Virtual Event, The Netherlands, October 25-29, 2020*, 2020, pp. 760–761, Association for Computing Machinery.

[19] J. L. Davis, A. Williams, M. W. Yang, "Algorithmic reparation," *Big Data and Society*, vol. 8, 2021, doi: 10.1177/20539517211044808.

[20] A. M. Antoniadi, Y. Du, Y. Guendouz, L. Wei, C. Mazo, B. A. Becker, C. Mooney, "Current challenges and future opportunities for xai in machine learning- based clinical decision support systems: A systematic review," *Applied Sciences (Switzerland)*, vol. 11, 6 2021, doi: 10.3390/app11115088.

[21] K. Sokol, "Fairness, accountability and transparency in artificial intelligence: A case study of logical predictive models," in *Proceedings of the 2019 Conference on AI, Ethics, and Society*, New York, NY, USA, 2019, p. 541–542, Association for Computing Machinery.

[22] K. Gade, S. C. Geyik, K. Kenthapadi, V. Mithal, A. Taly, "Explainable AI in industry: practical challenges and lessons learned: implications tutorial," in *Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, 2020, p. 699, Association for Computing Machinery.

[23] A. Stevens, P. Deruyck, Z. V. Veldhoven, J. Vanthienen, "Explainability and fairness in machine learning: Improve fair end-to-end lending for kiva," in *2020 IEEE Symposium Series on Computational Intelligence Canberra, Australia, December 1-4, 2020*, 2020, pp. 1241–1248, IEEE.

[24] R. Bellamy, K. Dey, M. Hind, S. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. Natesan Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. Varshney, Y. Zhang, "Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," *IBM Journal of Research and Development*, vol. PP, 09 2019, doi: 10.1147/JRD.2019.2942287.

[25] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. A. Specter, L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *5th IEEE International Conference on Data Science and Advanced Analytics, Turin, Italy, October 1-3, 2018*, 2018, pp. 80–89, IEEE.

[26] E. Mutlu, O. O. Garibay, "A quantum leap for fairness: Quantum bayesian approach for fair decision making," in *Human-Computer Interaction (HCI) International 2021 - Late Breaking Papers: Multimodality, eXtended Reality, and Artificial Intelligence*, 2021, pp. 489–499, Springer International Publishing.

[27] J. Stoyanovich, B. Howe, S. Abiteboul, G. Miklau, A. Sahuguet, G. Weikum, "Fides: Towards a platform for responsible data science," in *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, New York, NY, USA, 2017, Association for Computing Machinery.

[28] C. Addis, M. Kutar, "AI management an exploratory survey of the influence of GDPR and FAT principles," in *2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation, Leicester, United Kingdom, August 19-23, 2019*, 2019, pp. 342–347, IEEE.

[29] A. B. Arrieta, N. Díaz-Rodríguez, J. D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, Herrera, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020, doi: https://doi.org/10.1016/j.inffus.2019.12.012.

[30] V. Bhargava, M. Couceiro, A. Napoli, "Limeout: An ensemble approach to improve process fairness," in *European Conference on Machine Learning and Knowledge Discovery in Databases 2020 Workshops*, 2020, pp. 475–491, Springer International Publishing.

[31] S. Wachter, B. Mittelstadt, C. Russell, "Why fairness cannot be automated: Bridging the gap between eu non-discrimination law and ai," *Computer Law and Security Review*, vol. 41, 7 2021, doi: 10.1016/j.clsr.2021.105567.

[32] S. Abbasi-Sureshjani, R. Raumanns, B. E. J. Michels, Schouten, V. Cheplygina, "Risk of training diagnostic algorithms on data with demographic bias," in *Interpretable and Annotation-Efficient Learning for Medical Image Computing*, 2020, pp. 183–192, Springer International Publishing.

[33] M. DeCamp, C. Lindvall, "Latent bias and the implementation of artificial intelligence in medicine," *Journal of the American Medical Informatics Association*, vol. 27, pp. 2020–2023, 12 2020, doi: 10.1093/jamia/ocaa094.

[34] Y. Zheng, S. Wang, J. Zhao, "Equality of opportunity in travel behavior prediction with deep neural networks and discrete choice models," *Transportation Research Part C: Emerging Technologies*, vol. 132, p. 103410, 2021, doi: https://doi.org/10.1016/j.trc.2021.103410.

[35] V. V. Vesselinov, B. S. Alexandrov, D. O'Malley, "Nonnegative tensor factorization for contaminant source identification," *Journal of Contaminant Hydrology*, vol. 220, pp. 66–97, 2019, doi: https://doi.org/10.1016/j.jconhyd.2018.11.010.

[36] O. J. Akintande, "Algorithm fairness through data inclusion, participation, and reciprocity," in *Database Systems for Advanced Applications*, 2021, pp. 633–637, Springer International Publishing.

[37] S. Park, H. Ko, "Machine learning and law and economics: A preliminary overview," *Asian Journal of Law and Economics*, vol. 11, 8 2020, doi: 10.1515/ajle-2020-0034.

[38] F. Marcinkowski, K. Kieslich, C. Starke, M. Lünich, "Implications of AI (un-)fairness in higher education admissions: the effects of perceived AI (un-)fairness on exit, voice and organizational reputation," in *Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, 2020, pp. 122– 130, Association for Computing Machinery.

[39] D. Solans, F. Fabbri, C. Calsamiglia, C. Castillo, F. Bonchi, "Comparing equity and effectiveness of different algorithms in an application for

the room rental market," in *Proceedings of the 2021 Conference on AI, Ethics, and Society*, New York, NY, USA, 2021, p. 978–988, Association for Computing Machinery.

[40] S. Hajian, F. Bonchi, C. Castillo, "Algorithmic bias: From discrimination discovery to fairness- aware data mining," in *Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, 2016, pp. 2125–2126, Association for Computing Machinery.

[41] C. M. Madla, F. K. H. Gavins, H. A. Merchant, M. Orlu, S. Murdan, A. W. Basit, "Let's talk about sex: Differences in drug therapy in males and females," *Advanced Drug Delivery Reviews*, vol. 175, p. 113804, 2021, doi: https://doi.org/10.1016/j.addr.2021.05.014.

[42] G. Currie, K. E. Hawk, "Ethical and legal challenges of artificial intelligence in nuclear medicine," *Seminars in Nuclear Medicine*, vol. 51, pp. 120–125, 2021, doi: https://doi.org/10.1053/j.semnuclmed.2020.08.001.

[43] G. Starke, E. D. Clercq, B. S. Elger, "Towards a pragmatist dealing with algorithmic bias in medical machine learning," *Medicine, Health Care and Philosophy*, vol. 24, pp. 341–349, 9 2021, doi: 10.1007/s11019-021-10008-5.

[44] A. M. Fejerskov, "Algorithmic bias and the (false) promise of numbers," *Global Policy*, vol. 12, pp. 101– 103, 7 2021, doi: 10.1111/1758-5899.12915.

[45] C. Panigutti, A. Perotti, A. Panisson, P. Bajardi, D. Pedreschi, "Fairlens: Auditing black-box clinical decision support systems," *Information Processing & Management*, vol. 58, p. 102657, 2021, doi: https://doi.org/10.1016/j.ipm.2021.102657.

[46] S. Kino, Y.-T. Hsu, K. Shiba, Y.-S. Chien, C. Mita, I. Kawachi, A. Daoud, "A scoping review on the use of machine learning in research on social determinants of health: Trends and research prospects," *SSM-Population Health*, vol. 15, p. 100836, 2021, doi: https://doi.org/10.1016/j.ssmph.2021.100836.

[47] N. Norori, Q. Hu, F. M. Aellen, F. D. Faraci, A. Tzovara, "Addressing bias in big data and ai for health care: A call for open science," *Patterns*, vol. 2, p. 100347, 2021, doi: https://doi.org/10.1016/j.patter.2021.100347.

[48] L. Xu, "The dilemma and countermeasures of ai in educational application," *Pervasive Health: Pervasive Computing Technologies for Healthcare*, pp. 289–294, 12 2020, doi: 10.1145/3445815.3445863.

[49] W. Holmes, K. Porayska-Pomsta, K. Holstein, E. Sutherland, T. Baker, S. B. Shum, O. C. Santos, M. T. Rodrigo, M. Cukurova, I. I. Bittencourt, K. R. Koedinger, "Ethics of ai in education: Towards a community-wide framework," *International Journal of Artificial Intelligence in Education*, 2021, doi: 10.1007/s40593-021-00239-1.

[50] A. Peña, I. Serna, A. Morales, J. Fierrez, "Faircvtest demo: Understanding bias in multimodal learning with a testbed in fair automatic recruitment," *2020 International Conference on Multimodal Interaction*, pp. 760–761, 10 2020, doi: 10.1145/3382507.3421165.

[51] A. Ortega, J. Fierrez, A. Morales, Z. Wang, M. de la Cruz, C. L. Alonso, T. Ribeiro, "Symbolic ai for xai: Evaluating lfit inductive programming for explaining biases in machine learning," *Computers*, vol. 10, 11 2021, doi: 10.3390/computers10110154.

[52] S. K. Misra, S. Das, S. Gupta, S. K. Sharma, "Public policy and regulatory challenges of artificial intelligence (ai)," *Advances in Information and Communication Technology*, vol. 617, pp. 100–111, 2020, doi: 10.1007/978-3-030-64849-7_10.

[53] S. Pundhir, U. Ghose, V. Kumari, "Legitann: Neural network model with unbiased robustness," in *Proceedings of International Conference on Communication and Artificial Intelligence*, 2021, pp. 385– 397, Springer Singapore.

[54] S. Sharma, A. H. Gee, D. Paydarfar, J. Ghosh, "Fair- n: Fair and robust neural networks for structured data," in *Proceedings of the 2021 Conference on AI, Ethics, and Society*, New York, NY, USA, 2021, p. 946–955, Association for Computing Machinery.

[55] J. Kang, H. Tong, "Fair graph mining," in *The 30th International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*, 2021, pp. 4849–4852, Association for Computing Machinery.

[56] J. Bobadilla, R. Lara-Cabrera, Á. González- Prieto, F. Ortega, "DeepFair: Deep learning for improving fairness in recommender systems," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 6, p. 86, 2021, doi: 10.9781/ijimai.2020.11.001.

[57] N. Martinez-Martin, Z. Luo, A. Kaushal, E. Adeli, A. Haque, S. S. Kelly, S. Wieten, M. K. Cho, D. Magnus, L. Fei-Fei, K. Schulman, A. Milstein, "Ethical issues in using ambient intelligence in health-care settings," *The Lancet Digital Health*, vol. 3, pp. e115–e123, 2 2021, doi: 10.1016/S2589-7500(20)30275-2.

[58] S. K. Kane, A. Guo, M. R. Morris, "Sense and accessibility: Understanding people with physical disabilities' experiences with sensing systems," in *The 22nd International Conference on Computers and Accessibility, Virtual Event, Greece, October 26-28, 2020*, 2020, pp. 42:1–42:14, Association for Computing Machinery.

[59] A. Paviglianiti, E. Pasero, "VITAL-ECG: a de-bias algorithm embedded in a gender-immune device," in *2020 IEEE International Workshop on Metrology for Industry 4.0 & IoT, Roma, Italy, June 3-5, 2020*, 2020, pp. 314–318, IEEE.

[60] R. Mark, "Ethics of public use of ai and big data: The case of amsterdam's crowdedness project," *The ORBIT Journal*, vol. 2, no. 2, pp. 1–33, 2019, doi: https://doi.org/10.29297/orbit.v2i1.101.

[61] C. E. Kontokosta, B. Hong, "Bias in smart city governance: How socio-spatial disparities in 311 complaint behavior impact the fairness of data-driven decisions," *Sustainable Cities and Society*, vol. 64, p. 102503, 2021, doi: https://doi.org/10.1016/j.scs.2020.102503.

[62] V. D. Badal, C. Nebeker, K. Shinkawa, Y. Yamada, K. E. Rentscher, H.-C. Kim, E. E. Lee, "Do words matter? detecting social isolation and loneliness in older adults using natural language processing," *Frontiers in Psychiatry*, vol. 12, 11 2021, doi: 10.3389/fpsyt.2021.728732.

[63] B. Richardson, D. Prioleau, K. Alikhademi, J. E. Gilbert, "Public accountability: Understanding sentiments towards artificial intelligence across dispositional identities," in *IEEE International Symposium on Technology and Society, Tempe, AZ, USA, November 12-15, 2020*, 2020, pp. 489–496, IEEE.

[64] D. Muralidhar, "Examining religion bias in AI text generators," in *Conference on AI, Ethics, and Society, Virtual Event, USA, May 19-21, 2021*, 2021, pp. 273– 274, Association for Computing Machinery.

[65] A. J. Larrazabal, N. Nieto, V. Peterson, D. H. Milone, E. Ferrante, "Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 117, pp. 12592–12594, 6 2020, doi: 10.1073/pnas.1919012117.

[66] E. Puyol-Antón, B. Ruijsink, S. K. Piechnik, S. Neubauer, S. E. Petersen, R. Razavi, A. P. King, "Fairness in cardiac mr image analysis: An investigation of bias due to data imbalance in deep learning based segmentation," *Medical Image Computing and Computer Assisted Intervention*, vol. 12903 LNCS, pp. 413–423, 2021, doi: 10.1007/978- 3-030-87199-4_39.

[67] A. Rajabi, O. O. Garibay, "Towards fairness in ai: Addressing bias in data using gans," in *Human- Computer Interaction (HCI) International 2021 - Late Breaking Papers: Multimodality, eXtended Reality, and Artificial Intelligence*, Cham, 2021, pp. 509–518, Springer International Publishing.

[68] Y. Zhang, J. Sang, "Towards accuracy-fairness paradox: Adversarial example-based data augmentation for visual debiasing," in *Proceedings of the 28th International Conference on Multimedia*, New York, NY, USA, 2020, p. 4346–4354, Association for Computing Machinery.

[69] A. K. Singh, A. Blanco-Justicia, J. Domingo-Ferrer, D. Sánchez, D. Rebollo-Monedero, "Fair detection of poisoning attacks in federated learning," in *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence*, 2020, pp. 224–229, IEEE.

[70] C. Fan, M. Esparza, J. Dargin, F. Wu, B. Oztekin, A. Mostafavi, "Spatial biases in crowdsourced data: Social media content attention concentrates on populous areas in disasters," *Computers, Environment and Urban Systems*, vol. 83, p. 101514, 2020, doi: https://doi.org/10.1016/j.compenvurbsys.2020.101514.

[71] R. Chiong, Z. Fan, Z. Hu, F. Chiong, "Using an improved relative error support vector machine for body fat prediction," *Computer Methods and Programs in Biomedicine*, vol. 198, p. 105749, 2021, doi: https://doi.org/10.1016/j.cmpb.2020.105749.

[72] W. Lee, H. Ko, J. Byun, T. Yoon, J. Lee, "Fair clustering with fair correspondence distribution," *Information Sciences*, vol. 581, pp. 155–178, 2021, doi: https://doi.org/10.1016/j.ins.2021.09.010.

[73] W. Zhang, A. Bifet, X. Zhang, J. C. Weiss, W. Nejdl, "Farf: A fair and adaptive random forests classifier," in *Advances in Knowledge Discovery and Data Mining*, 2021, pp. 245–256, Springer International Publishing.

[74] C. G. Harris, "Mitigating cognitive biases in machine learning algorithms for decision making," in *Companion Proceedings of the Web Conference*

2020, New York, NY, USA, 2020, p. 775–781, Association for Computing Machinery.

[75] M. A. U. Alam, "Ai-fairness towards activity recognition of older adults," *Pervasive Health: Pervasive Computing Technologies for Healthcare*, pp. 108–117, 12 2020, doi: 10.1145/3448891.3448943.

[76] Y. Zhang, A. Ramesh, "Learning fairness-aware relational structures," *Frontiers in Artificial Intelligence and Applications*, vol. 325, pp. 2543–2550, 8 2020, doi: 10.3233/FAIA200389.

[77] S. Ahmed, S. A. Athyaab, S. A. Muqtadeer, "Attenuation of human bias in artificial intelligence: An exploratory approach," in *2021 6th International Conference on Inventive Computation Technologies*, 2021, pp. 557–563, IEEE.

[78] Y. Hou, H. Hong, Z. Sun, D. Xu, Z. Zeng, "The control method of twin delayed deep deterministic policy gradient with rebirth mechanism to multi-dof manipulator," *Electronics (Switzerland)*, vol. 10, 4 2021, doi: 10.3390/electronics10070870.

[79] K. Xivuri, H. Twinomurinzi, "A systematic review of fairness in artificial intelligence algorithms," in *Responsible AI and Analytics for an Ethical and Inclusive Digitized Society*, Cham, 2021, pp. 271–284, Springer International Publishing.

[80] S. Sharma, Y. Zhang, J. M. Ríos Aliaga, D. Bouneffouf, V. Muthusamy, K. R. Varshney, "Data augmentation for discrimination prevention and bias disambiguation," in *Proceedings of the Conference on AI, Ethics, and Society*, New York, NY, USA, 2020, p. 358–364, Association for Computing Machinery.

[81] A. Pandey, A. Caliskan, "Disparate impact of artificial intelligence bias in ridehailing economy's price discrimination algorithms," in *Conference on AI, Ethics, and Society, Virtual Event, USA, May 19-21, 2021*, 2021, pp. 822–833, Association for Computing Machinery.

[82] S. Udeshi, P. Arora, S. Chattopadhyay, "Automated directed fairness testing," in *Proceedings of the 33rd International Conference on Automated Software Engineering*, New York, NY, USA, 2018, p. 98–108, Association for Computing Machinery.

[83] N. V. Berkel, J. Goncalves, D. Hettiachchi, S. Wijenayake, R. M. Kelly, V. Kostakos, "Crowdsourcing perceptions of fair predictors for machine learning: A recidivism case study," *Proceedings of the Association for Computing Machinery on Human-Computer Interaction*, vol. 3, 11 2019, doi: 10.1145/3359130.

[84] W. Fleisher, "What's fair about individual fairness?," in *Proceedings of the 2021 Conference on AI, Ethics, and Society*, New York, NY, USA, 2021, p. 480–490, Association for Computing Machinery.

[85] O. Aka, K. Burke, A. Bauerle, C. Greer, M. Mitchell, "Measuring model biases in the absence of ground truth," in *Proceedings of the 2021 Conference on AI, Ethics, and Society*, New York, NY, USA, 2021, p. 327–335, Association for Computing Machinery.

[86] S. Verma, J. Rubin, "Fairness definitions explained," in *Proceedings of the International Workshop on Software Fairness*, New York, NY, USA, 2018, p. 1–7, Association for Computing Machinery.

[87] D. Fan, Y. Wu, X. Li, "On the fairness of swarm learning in skin lesion classification," *Clinical Image-Based Procedures, Distributed and Collaborative Learning, Artificial Intelligence for Combating COVID-19 and Secure and Privacy-Preserving Machine Learning"*, p. 120–129, 2021, doi: 10.1007/978-3-030-90874-4_12.

[88] T. Speicher, H. Heidari, N. Grgic-Hlaca, K. P. Gummadi, A. Singla, A. Weller, M. B. Zafar, "A unified approach to quantifying algorithmic unfairness," *Proceedings of the 24th International Conference on Knowledge Discovery and Data Mining*, jul 2018, doi: 10.1145/3219819.3220046.

[89] L. Liang, D. E. Acuna, "Artificial mental phenomena: Psychophysics as a framework to detect perception biases in ai models," *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 403–412, 1 2020, doi: 10.1145/3351095.3375623.

[90] P. Saleiro, B. Kuester, A. Stevens, A. Anisfeld, L. Hinkson, J. London, R. Ghani, "Aequitas: A bias and fairness audit toolkit," *CoRR*, vol. abs/1811.05577, 2018.

[91] H. J. P. Weerts, M. Dudík, R. Edgar, A. Jalali, R. Lutz, M. Madaio, "Fairlearn: Assessing and improving fairness of AI systems," *Computing Research Repository*, vol. abs/2303.16626, 2023, doi: 10.48550/arXiv.2303.16626.

[92] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, J. Wilson, "The what-if tool: Interactive probing of machine learning models," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 56–65, 2020, doi: 10.1109/TVCG.2019.2934619.

[93] C. Wilson, A. Ghosh, S. Jiang, A. Mislove, L. Baker, J. Szary, K. Trindel, F. Polli, "Building and auditing fair algorithms: A case study in candidate screening," in *Proceedings of the 2021 Conference on Fairness, Accountability, and Transparency*, New York, NY, USA, 2021, p. 666–677, Association for Computing Machinery.

[94] A. Pérez-Suay, V. Laparra, G. Mateo-García, J. Muñoz- Marí, L. Gómez-Chova, G. Camps-Valls, "Fair kernel learning," in *Machine Learning and Knowledge Discovery in Databases*, 2017, pp. 339–355, Springer International Publishing.

[95] P. Smith, K. Ricanek, "Mitigating algorithmic bias: Evolving an augmentation policy that is non-biasing," in *2020 IEEE Winter Applications of Computer Vision Workshops*, 2020, pp. 90–97, IEEE.

[96] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, S. Venkatasubramanian, "Certifying and removing disparate impact," in *Proceedings of the 21th International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2015, p. 259–268, Association for Computing Machinery.

[97] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, "Generative adversarial networks," *Advances in Neural Information Processing Systems*, vol. 3, 06 2014, doi: 10.1145/3422622.

[98] B. H. Zhang, B. Lemoine, M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *Proceedings of the 2018 Conference on AI, Ethics, and Society*, New York, NY, USA, 2018, p. 335–340, Association for Computing Machinery.

[99] C. Weber, "Engineering bias in ai," *IEEE Pulse*, vol. 10, pp. 15–17, 1 2019, doi: 10.1109/MPULS.2018.2885857.

[100] N. T. Lee, "Detecting racial bias in algorithms and machine learning," *Journal of Information, Communication and Ethics in Society*, vol. 16, pp. 252– 260, 11 2018, doi: 10.1108/JICES-06-2018-0056.

[101] N. McDonald, S. Pan, "Intersectional ai: A study of how information science students think about ethics and their impact," *Proceedings of the Association for Computing Machinery on Human-Computer Interaction*, vol. 4, 10 2020, doi: 10.1145/3415218.

[102] P. K. Bharti, T. Ghosal, M. Agrawal, A. Ekbal, "How confident was your reviewer? estimating reviewer confidence from peer review texts," in *Document Analysis Systems - 15th International Workshop, 2022, La Rochelle, France, May 22-25, 2022, Proceedings*, vol. 13237 of *Lecture Notes in Computer Science*, 2022, pp. 126–139, Springer.

Rubén González Sendino

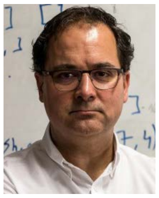
Rubén González completed his Master's degree in Artificial Intelligence from the Universidad Politécnica de Madrid in 2017. He is currently pursuing a PhD at the same institution. Over the years, he has gained extensive experience in innovation departments and has actively collaborated on European Horizon 2020 projects. He possesses expertise in deep neural networks, natural language processing, causal algorithms, and computer vision. Beyond the technical intricacies, he is deeply committed to explicability and responsible artificial intelligence.

Emilio Serrano

Emilio Serrano received the M.Sc. degree in computer science (2006) and the Ph.D. degree, with European mention and Extraordinary Ph.D. Award in artificial intelligence (2011), from the University of Murcia, Spain. He has also been a Visiting Researcher with The University of Edinburgh, the University of Oxford, and the National Institute of Informatics in Tokyo. He is currently an Associate Professor with the Department of Artificial Intelligence, Universidad Politécnica de Madrid (UPM). His main research line is the Social and Explainable Artificial Intelligence for Smart Cities. His scientific production includes more than 80 publications. He lectures deep learning for natural language processing and social network analysis among other courses. He has been principal investigator in four educational innovation projects in data science, participated in several European and National funding programs (6 European projects), and supervised two Ph.D theses.

Javier Bajo

Dr. Javier Bajo, full professor at the Department of Artificial Intelligence, Computer Science School at Universidad Politécnica de Madrid (UPM), holds (since 03/05/2019) the position of Director of the UPM AI.nnovation Space Research Center in Artificial Intelligence. He was Director of the Department of Artificial Intelligence (20/05/2016-19/10/2017) at UPM, Secretary of the PhD in Artificial Intelligence at UPM (23/06/2016-19/10/2017) and Coordinator of the Research Master in Artificial Intelligence at UPM (18/02/2013 - 20/05/2016). He also holds the position of Director of the Data Center at the Pontifical University of Salamanca (13-10/2010 - 08-11-2012), with 21 employees. His main lines of research are Social Computing and Artificial and Hybrid Societies; Intelligent Agents and Multiagent Systems, Ambient Intelligence, Machine Learning. He has supervised 11 Ph.D thesis, participated in more than 50 research projects (in most of them as principal investigator) and published more than 300 articles in recognized journals (81 JCR papers) and conferences. His h-index is 39. He is founder of the PAAMS series of conferences and is an IEEE, ACM and ISIF member.



Paulo Novais

Dr. Paulo Novais is a Full Professor of Computer Science at the Department of Informatics, the School of Engineering, the University of Minho (Portugal) and a researcher at the ALGORITMI Centre in which he is the leader of the research group ISLab - Synthetic Intelligence lab. He is the director of the PhD Program in Informatics and co-founder and Deputy Director of the Master in Law and Informatics at the University of Minho. His main research objective is to make systems a little more smart, reliable and sensitive to human presence and interaction. He is the coordinator of the Portuguese Intelligent Systems Associate Laboratory (LASI). He has supervised 20 PhD thesis, participated in several research projects sponsored by Portuguese and European public and private Institutions, and published more than 350 articles in recognized journals and conferences.

Analysis of Gender Differences in Facial Expression Recognition Based on Deep Learning Using Explainable Artificial Intelligence

Cristina Manresa-Yee*, Silvia Ramis, José M. Buades

Universitat de les Illes Balears, Group of Computer Graphics, Computer Vision and IA. Maths and Computer Science Department, Palma (Spain)

* Corresponding author: cristina.manresa@uib.es

Received 29 July 2022 | Accepted 3 March 2023 | Published 12 April 2023

LA UNIVERSIDAD
EN INTERNET

Abstract

Potential uses of automated Facial Expression Recognition (FER) cover a wide range of applications such as customer behavior analysis, healthcare applications or providing personalized services. Data for machine learning play a fundamental role, therefore, understanding the relevancy of the data in the outcomes is of utmost importance. In this work we present a study on how gender influences the learning of a FER system. We analyze with Explainable Artificial intelligence (XAI) techniques how gender contributes to the learning and assess which facial expressions are more similar regarding face regions that impact on the classification. Results show that there exist common regions in some expressions both for females and males with different intensities (e.g. happiness); however, there are other expressions like disgust, where important face regions differ. The insights of this work will help improving FER systems and understand the source of any inequality.

Keywords

Explainable AI, Facial Expression Recognition, Gender Differences, Explainable Artificial Intelligence, XAI.

DOI: 10.9781/ijimai.2023.04.003

I. Introduction

That men's and women's facial expressions differ is a well-known fact. Gender differences have been examined subjectively but also on physiological measures such as facial electromyography (EMG) or observable Action Units (AUs). Studies show differences in frequency [1]–[3], increased attention [4] or expressiveness specially in the expressions of happiness, fear, anger or disgust [1], [5], [6]. Regarding smiling, Dimberg and Lundquist [3] found that women evoked more facial muscle activity in response to happy faces. This result is consistent with earlier works [7]. Regarding fear, evidence shows that females were more facially expressive when presented with fear-relevant images (e.g. angry faces or snakes) with an increase of the activity on the corrugator supercilia [6]. Women also experience disgust with more intensity than men [4], however, anger is less likely to be displayed by females [8].

Six basic facial expressions -happiness, sadness, anger, surprise, disgust and fear- are recognized across different cultures [9]. Descriptions have been made about the face muscles involved in forming those expressions. The Facial Action Coding System (FACS) [10] describes anatomically all visually discernible facial movement by defining Action units (AUs), which are the actions of individual muscles or groups of muscles. Observing and coding a selection of AUs, Emotion FACS (EMFACS), humans can identify prototypical facial expressions that have been found to suggest certain emotions. Fan, Lan and Li [11] analyzed two of these AUS, the AU6 (cheek raiser) and AU12 (lip corner puller) related to the smiling (happiness expression) and found that females were generally more expressive and presented a higher intensity value for AU12 (bigger smile) than males. McDuff et al. [1] also analyzed AUs (AU1, AU2, AU4, AU12 and AU15) to study gender differences. Their results found that women smiled more, and they presented more significantly inner brow raise actions, which are related with fear and sadness.

Houstis and Kiliaridis [12] did not use AUs, but they analyzed a set of facial distances when posing a rest pose, a lip pucker, and a posed smile. Their findings regarding gender, found that males had a vertical upward component more pronounced both in the posed smile and the lip pucker, while females had a more pronounced horizontal component in the posed smile.

There is extensive research in analyzing gender differences both for recognition (perceiver) [13]–[15] and generation (expresser) of facial expressions [16], [17]. However, in the case of current deep learning developments for automatic FER [18], due to their black box nature, it makes it difficult to assess the gender differences in recognition beyond comparisons of for example the accuracy rates or bias [19]–[22]. However, a deeper understanding on how men and women

Please cite this article as:
C. Manresa-Yee, S. Ramis, J. M. Buades. Analysis of Gender Differences in Facial Expression Recognition Based on Deep Learning Using Explainable Artificial Intelligence, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 1, pp. 18-27, 2024, http://dx.doi.org/10.9781/ijimai.2023.04.003

images contribute to the learning of the models could help improving the models or understanding the misclassifications. Further, research in this area is usually based on existing datasets which are not always gender balanced [18] and we even find datasets, such as JAFFE [23] , with only one gender (10 Japanese female subjects).

Explainable Artificial Intelligence (XAI) techniques can provide further information on the internal working on these models and make them more transparent. Although they do not include a gender perspective, we find examples applying these techniques in automated FER to understand automatic emotional annotation [24], to improve transfer learning [25], [26] or to understand the influential face regions in the classification [27], [28]. Heimerl et al. [24] included XAI techniques in their emotional behavior annotation tool addressed to non-expert users. Humans assisted the automatic labeling -only for four out of Ekman's six basic emotions: happiness, sadness, anger and disgust- aided by confidence values of the predicted annotation, as well as visual explanations using XAI (LIME [29] , INNvestigate [30]). Schiller et al. [25] presented saliency maps to identify the most relevant face regions used for the face recognition. The saliency maps were generated by Layer-wise Relevance Propagations (LRP) [31] and by eye-tracking. Then, they evaluated both and transferred that knowledge by hiding the non-relevant information to speed up the training of the neural network in a new domain.

Weitz et al. [27] investigated a Convolutional Neural Network (CNN) trained to distinguish facial expressions of pain, happiness, and disgust. They applied two XAI methods: LRP and LIME. They observed that the CNN did not exclusively look at the face but also to the background of the image. Regarding pain, Prajod et al. [26] also presented a study on the effects of transfer learning for automatic FER for emotions to pain. They applied LRP saliency maps to visually compare and understand the most influent regions for the classification, both for emotion recognition and for pain recognition, and related those regions with AUs. The results showed that specific AUs related to the facial expressions of contempt and surprise were not relevant for pain recognition.

With a gender perspective in mind, in our previous exploratory study, we could sense differences in the learning [32] which motivated us to analyze more thoroughly the impact of gender in automated FER. Further, the interest in this field is due to the multiple and varied domains that can benefit from FER such as diagnosis and treatment of psychiatric illness [33], marketing psychology applications [34] or human computer interfaces [35]. Therefore, by studying the influence of gender differences in FER training, we can improve our understanding of which face regions are important and consider this knowledge to contribute with better models which will impact the applications based on FER.

The work is organized as follows: Section 2 describes the material used and the procedure followed for the study. Section 3 presents and discusses the main results regarding performance, and gender differences and similarities in the important face regions considered by the model. Finally, the main contributions and future line works are presented in the last Section.

II. Materials and Methods

This section contains detailed description of data, data preprocessing and augmenting, the XAI approach used to understand the internal working of the model and the procedure followed.

A. Dataset

We train our FER model on the AffectNet dataset [36], a well-known public dataset and widely used in FER. AffectNet comprises more than one million still images of facial expressions in the wild and covers both categorical and dimensional affect models. About half (approximately 440K images) of the images are manually annotated as one of Ekman's basic emotions [37] (anger, fear, disgust, sadness, surprise, happiness, contempt and neutral). Further, the dataset presents issues such as duplicates or non-face images because it was built through web-scraping. In order to study gender differences, we manually labelled images (Female - Male) and selected a similar number of images regarding the gender label and the facial expression (see Fig. 1). Although there are a few datasets with gender labels and facial expressions (e.g. RAFDB with around 30K images [38]), we selected AffectNet due to its size and wide application in FER. It is important to highlight that formally both sex or gender should be informed or self-reported, but on web-scrapped datasets that information is not available. According to the World Health Organization (WHO), sex refers to "the different biological and physiological characteristics of males and females" and gender refers to "the socially constructed characteristics of women and men – such as norms, roles and relationships of and between groups of women and men" [39]. The view of gender used in this paper is binary. When we refer to a particular gender, we are assuming this gender based on the visual characteristics of the individual in the image. Additionally, as Chen and Joo [40] identified, human-generated annotations in FER datasets can include biases (e.g. annotation biases between genders, especially when it comes to the happy and angry expressions). Therefore, this study is limited due to these reasons.

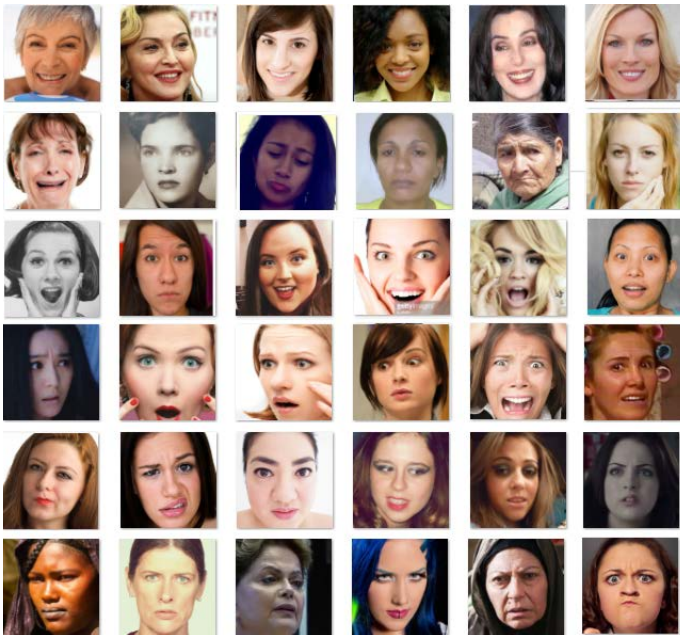


Fig. 1. Excerpt from the dataset of females' images for each expression.

The dataset used in the experimentation is comprised by an initial subset of 19044 images (1587 images x six expression x two gender) randomly selected where duplicated images, non-face images and images of individuals difficult to identify their gender by observation (e.g. babies or androgynous faces) were not considered. To meet the balanced dataset requirement, the subset was chosen considering the maximum number of manually labelled images per gender and facial expression, which was limited by the expression of Disgust. The Disgust expression counts with 4303 images in AffectNet, but 553 are duplicates, 290 images were undetermined and 1873 correspond to males and 1587 correspond to females. Therefore, the maximum number of images per gender and facial expression is 1587. Table I describes quantitatively the dataset in terms of expression and gender.

TABLE I. Number of Images in Terms of Facial Expression and Gender From the Original Dataset to the Training and Testing Datasets. In Gray the Expressions Used in This Study. Dupl.: Duplicates. Und: Undetermined. F: Female. M: Male.

			Gender label			Selected		Pre-processing Face detected			Female datasets		Male datasets	
	#	Dupl.	F	M	Und.	F	M	F	M	|F-M|	Test	Train	Test	Train
Neutral	75374	5093												
Happiness	134915	8430	2458	1590	581	1587	1587	1254	1182	72	253	1001	238	944
Sadness	25959	3425	1588	1691	613	1587	1587	1130	1038	92	227	903	209	829
Surprise	14590	1183	1588	1670	681	1587	1587	1183	1058	125	237	946	214	844
Fear	6878	1082	1608	1588	553	1587	1587	1150	1023	127	235	915	206	817
Disgust	4303	553	1587	1873	290	1587	1587	1230	1093	137	243	987	218	875
Anger	25382	3169	1588	3944	475	1587	1587	1233	1123	110	246	987	225	898
Contempt	4250	315												
None	33588	2342												
Uncertain	12145	943												
Non-face	82915													
Non-labeled	6999													
Total	427298	26535	10417	12356	3193	9522	9522	7180	6517		1441	5739	1310	5207

B. Pre-Processing and Data Augmentation

Images are pre-processed and augmented before the training. The pre-processing steps carried out are face detection, face alignment and cropping. To detect the face, we apply the *a contrario* framework proposed by Lisani et al. [41]. For its alignment, we initially detect the eyes using the 68 facial landmarks proposed by Sagonas et al. [42]. We find the geometric centroid of each eye from these landmarks and compute the distance between them to draw a straight line and calculate the rotation angle. This angle is then used to align the eyes horizontally. Finally, we crop the face and resize it to 224x224 pixels to feed it into the network as input.

The pre-processing steps discarded images (see Table I) when the algorithm did not detect the face (e.g. the face was not completely visible or it was a side face).

Finally, to increase the number of images to train and add diversity, we augmented data by modifying lighting and appearance [43]. We use the gamma correction technique (see Eq. 1) to modify the lighting conditions, with four gamma values ($\gamma = 0.5$, $\gamma = 1.0$, $\gamma = 1.5$ and $\gamma = 2.0$).

$$y = \left(\frac{x}{255}\right)^{\frac{1}{\gamma}} \cdot 255 \tag{1}$$

where x is the original image, y is the new image and γ is the value modified to change the illumination. To modify the appearance, we apply four-pixel translations of the image in both axes.

C. Explainability Approach

There exist a varied number of XAI techniques [44] for explaining models and data both at global and local levels. In this work, to analyze visually the outcome of the system, we use LIME, Local Interpretable Model-agnostic Explanation [29]. LIME can be applied on any classifier and offers locally faithful explanations of the instance being explained. When LIME is applied to images, explanations are the parts of the image that are most positive towards a certain class. We present a novel strategy to use LIME to acquire global knowledge based on instance-level information in this context. To study the differences of the model for female and male globally, we merge all LIME masks obtained from the same training set, test set and expression in an average heatmap.

Instead of computing the heatmap on the input space of the network (a 224 x 224 matrix), it is more relevant to compute it on the face representation space, that is, the parts of the face that are more relevant to identify one or another expression, regardless the orientation, translation or scaling of the image.

To compute the heatmap in the face space, we normalize all images with LIME applied to make the points of interest coincide. Faces are transformed so that the landmarks coincide with the normal form (see Fig. 2).

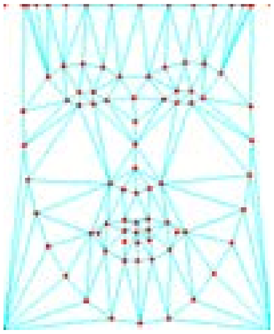
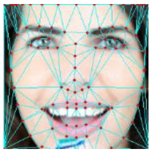
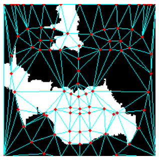
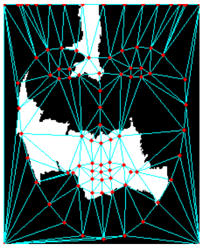
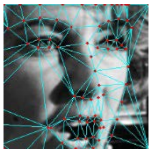
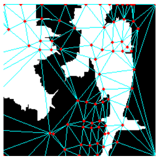
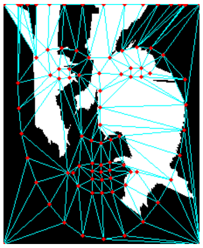
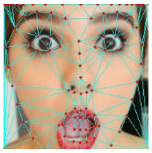
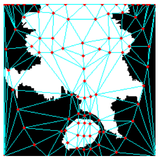
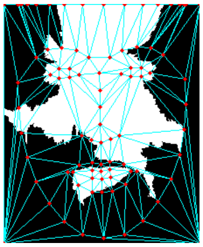
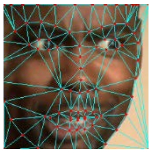
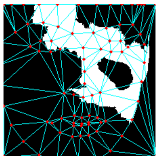
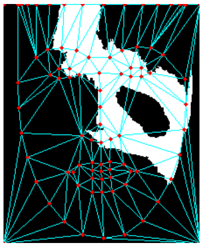
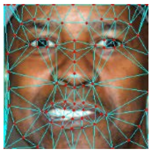
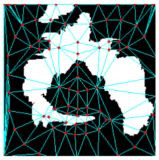
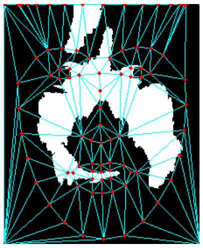
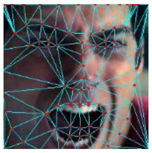
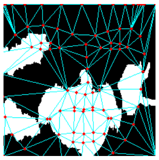
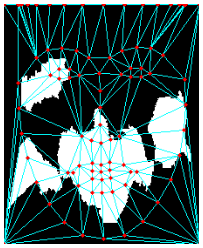


Fig. 2 Normalized location of the landmarks.

In Table II, we show the result of the process with six sample images, one for each expression starting from the original image: we identify the landmarks and compute the triangularization, then we apply LIME and superimpose the landmarks and triangularization, and finally we normalize this last image to the normal form. The detail of the process is described in the Algorithm 1.

```
Algorithm 1: Computing the normalized LIME image
procedure GetNormalizeLIME ($img$)                                   ▷ original image
    $black\_image$ ⟵ create black image with $img$ size
    $L'$ ⟵ landmarks($black\_image$)                                 ▷ 68 normalized landmarks
    $L'$ ⟵$L'$ ∪ 17 top points ∪ four corners                        ▷ 89 landmarks
    $L$ ⟵ landmarks ($img$)
    $L$ ⟵$L$ ∪ 17 top points ∪ four corners                          ▷ 89 landmarks
    $lime\_img$ ⟵ $lime$ ($img$)                                     ▷ Compute LIME for original image
    $tri$ ⟵ $delauny$ ($L$)                                          ▷ triangularization
    $norm\_lime\_img$ ⟵ empty image (224×275)                        ▷ Create empty image
    for each pixel coordinate $p' \in norm\_lime\_img$ do
        $(v'_i, v'_j, v'_k)$ ⟵ triangle from $L'$ that contains $p'$ using $tri$ triangles
        $(v_i, v_j, v'_k)$ ⟵ triangle from $L$ that match $(v'_i, v'_j, v'_k)$ in $L'$
        $(c'_i, c'_j, c'_k)$ ⟵ p^' coordinates as lineal combination of $(v'_i, v'_j, v'_k)$
        $p$ ⟵ lineal combination (barycentric coordinates) of $(v_i, v_j, v_k)$
            using $(c'_i, c'_j, c'_k)$ scalars
        $norm\_lime\_img\,[p'] = lime\_img\,[p]$
    return $norm\_lime\_img$                                         ▷ LIME image normalized
```

TABLE II. Normalization Process of the Images With LIME Applied to Merge the Important Regions for the Model

Exp.	Original image with landmarks and triangulation	LIME with landmarks and triangulation	LIME transformed to normal face with landmarks and triangulation
Happiness			
Sadness			
Surprise			
Fear			
Disgust			
Anger			

The merged heatmap is the average of all normalized LIME images that belong to the same training dataset, test dataset and facial expression. Then, we calculate distances between all the generated heatmaps, which is computed as one minus normalized correlation. Distances generate a symmetric matrix that is used to cluster similar heatmaps applying the Ward's variance minimization method [45]. Finally, we visualize the dendrogram generated by clustering to analyze the arrangement of the clusters produced by the models.

D. Procedure

We prepare three training and testing datasets: (1) a mixed dataset including a relatively gender-balanced number of images per facial expression; (2) a dataset with only images of females and (3) a dataset with only images of males.

Deep Convolutional Neural Networks (CNNs) have proved to be effective in numerous computer vision tasks [46], therefore, we use them to classify six facial expressions: anger, disgust, fear, happiness, sadness and surprise. In particular, we design our network based on the Inception v3 architecture [47]. The fully connected layer of Inception V3 network is replaced by a Global Average Pooling layer [48] and the softmax layer is modified to train the six classes (anger, sadness, fear, surprise, happiness, and disgust). Table III describes the hyper parameters of the network. The model is trained on a different dataset (mixed, only-female and only-male) and fine-tuned on imagenet [46]. We perform stratified 5-fold cross-validation, since these values have shown empirically to yield test error rate estimates that suffer neither from excessively high bias nor from very high variance [49], and report the mean classification accuracy. We highlight that the aim of the study is to analyze the influence of male and female datasets in the training, without focusing on improving the accuracy of the model. Then, the model is tested on all testing datasets (mixed, only-female and only-male) (see Table I for the details of number of images for each expression in each model). The different combinations of training and testing datasets used in this study are identified in Table IV.

TABLE III. Hyper-Parameters Used in the Inception V3 Network

Parameters	
Weights (pre-trained model)	Imagenet
Learning Rate	$lr = 10^{-4}$
Optimization algorithm	Adam [50]
Batch Size on training set	128
Batch Size on validation set	32
Epochs	1

TABLE IV. IDs of Combinations of Training and Testing Sets

		Testing		
		Mixed	Female	Male
Training	Mixed	MI-MI	MI-FE	MI-MA
	Female	FE-MI	FE-FE	FE-MA
	Male	MA-MI	MA-FE	MA-MA

To study differences between male and female, we apply the LIME merging procedure aforementioned to create the heatmaps to observe the face regions that are important for the model to classify images into a facial expression class. We build 36 heatmaps (3 training datasets (mixed, female and male) x 2 testing sets (female and male) x 6 expressions). In this case, LIME is configured to show the 5 most important features for the classification.

III. Results

In this section we present the accuracy obtained and the gender differences observed regarding the face regions that influence the recognition.

A. Accuracy

The training done with the mixed dataset achieves the best results in all cases (with accuracy around 53%-55%). Although we did not focus on improving the results, the accuracy is similar to other works that used AffectNet: Wang et al. [51] compiled several state-of-the-art methods on AffectNet with accuracies ranging from 47% to 60.23% for 7 or 8 expressions classification; Ngo et al. [52] achieved accuracies ranging from 46.07% to 60.7% using SE-ResNet-50 with different loss functions and classifying 8 expressions and Yen and Li [53] tested different architectures (ResNet-50, Xception, EfficientNet-B0, Inception, and DenseNet-121) with 8 expressions achieving accuracies ranging from 54% to 58% with class weight and data augmentation. Similar results are achieved both with the unbalanced trainings (male and female training) when tested with the mixed dataset (MA-MI, FE-MI). However, results decrease when testing with the other gender dataset, especially the training with female tested with male (FE-MA) (see Table V).

TABLE V. Mean Class-Wise Percentage Accuracy of the Models, Broken Down by Datasets

		Testing dataset		
		Mixed	**Male**	**Female**
Training dataset	**Mixed**	53.83	52.73	54.84
	Male	47.12	47.80	46.51
	Female	47.61	42.86	51.98

Observing the confusion matrices (see Tables VI, VII and VIII), happiness is the best recognized expression for all training and testing datasets, except for FE-MA (female training dataset and male testing dataset). Although accuracy is high both for the mixed and male trainings (above 75%) with all testing datasets, female training dataset values are around 70% only with the female testing datasets and achieves lower values with the other testing datasets (lower than 56%). It is noteworthy that the higher values testing with females are achieved both with the male (MA-FE) and mixed (MI-FE) training datasets, which might be because of the higher expressiveness of females when smiling [7].

In the case of mixed training, there is a similar behavior both for the testing with male and female (MI-MA, MI-FE), obtaining the worse classifications for the fear and anger expressions. Fear is highly misclassified with surprise, and anger with disgust and sadness. Even humans have difficulties identifying facial expressions such as disgust and anger [25].

In general, both surprise and anger are not well recognized when training with males but recognizing anger for females achieves only a 25% of accuracy (MA-FE). When training with females, surprise and fear are low recognized both for the mixed and male testing (FE-MI and FE-MA) and for female testing (FE-FE), fear and sadness present the lowest accuracy.

In FE-MA (female training dataset and male testing dataset), expressions tend to be classified as angry. That means that from the total of classifications, anger is the expression mostly selected by the CNN (see last row of each confusion matrix in Table VII). On the contrary, when using the MA-FE (male training dataset and female testing dataset), expressions tend to be classified into the happiness expression (see last row of each confusion matrix in Table VIII). Lastly, the mixed training dataset tends not to classify expressions into the fear class (see last row of each confusion matrix in Table VII and VIII).

TABLE VI. Confusion Matrix (Tested With Mixed Dataset, Trained With All Datasets). Values Are Expressed as Percentages. Last Row Is the Sum of All

Train/Test		**Mix**					
		Ha	Sa	Su	Fe	Di	An
Mix	Ha	79.07	4.43	6.29	0.86	5.58	3.78
	Sa	5.35	50.44	9.98	4.66	12.85	16.72
	Su	8.25	8.25	54.63	13.70	6.82	8.35
	Fe	3.27	10.83	30.32	36.89	8.87	9.81
	Di	6.97	11.24	6.49	4.21	54.03	17.06
	An	4.04	16.78	9.37	3.78	18.09	47.94
		107.0	102.0	117.1	64.1	106.3	103.7
Ma	Ha	78.19	5.30	5.46	2.67	6.41	1.97
	Sa	9.50	41.58	10.68	11.27	14.52	12.43
	Su	9.05	10.89	37.75	29.10	7.76	5.45
	Fe	5.01	10.28	22.05	47.13	9.20	6.32
	Di	12.10	14.24	6.21	9.63	45.28	12.54
	An	8.57	19.83	8.74	10.41	19.62	32.83
		122.4	102.1	90.9	110.2	102.8	71.6
Fe	Ha	55.51	8.95	7.23	3.16	17.07	8.08
	Sa	1.48	43.50	8.96	7.20	16.67	22.18
	Su	4.64	10.13	44.94	20.99	7.58	11.73
	Fe	1.57	10.57	24.64	40.35	9.25	13.63
	Di	3.05	12.45	5.99	6.51	50.15	21.85
	An	1.44	16.77	6.28	6.50	17.79	51.22
		67.7	102.4	98.0	84.7	118.5	128.7

TABLE VII. Confusion Matrix (Tested With Male Dataset, Trained With All Datasets). Values Are Expressed as Percentages. Last Row Is the Sum of All

Train/Test		**Male**					
		Ha	Sa	Su	Fe	Di	An
Mix	Ha	75.64	5.58	5.84	1.10	6.52	5.33
	Sa	5.20	52.59	7.92	3.37	12.80	18.12
	Su	6.72	11.24	51.26	15.20	6.51	9.07
	Fe	3.12	11.24	28.64	36.48	9.98	10.54
	Di	6.50	12.45	6.49	4.11	50.96	19.49
	An	3.83	18.21	7.91	4.28	16.31	49.45
		101.0	111.3	108.1	64.6	103.1	112.0
Ma	Ha	75.79	5.92	4.74	1.95	8.89	2.71
	Sa	8.68	41.96	8.87	8.59	15.02	16.88
	Su	8.23	12.00	35.77	26.70	8.90	8.41
	Fe	4.88	9.88	21.13	45.84	10.48	7.79
	Di	9.34	14.26	5.58	8.22	46.06	16.54
	An	5.53	17.20	6.67	9.11	20.07	41.42
		112.5	101.2	82.8	100.4	109.4	93.8
Fe	Ha	39.86	13.11	5.33	4.32	24.62	12.77
	Sa	0.58	43.63	4.64	6.36	20.12	24.68
	Su	2.65	13.88	32.14	25.89	9.45	15.99
	Fe	1.27	11.43	20.45	40.02	9.88	16.94
	Di	1.19	14.09	4.94	6.32	47.31	26.15
	An	0.44	18.44	2.94	6.96	17.01	54.21
		46.0	114.6	70.4	89.9	128.4	150.7

TABLE VIII. Confusion Matrix (Tested With Female Dataset, Trained With All Datasets). Values Are Expressed as Percentages. Last Row Is the Sum of All

Train/Test		Female					
		Ha	Sa	Su	Fe	Di	An
Mix	Ha	82.30	3.34	6.71	0.64	4.70	2.32
	Sa	5.50	48.46	11.89	5.84	12.89	15.42
	Su	9.63	5.58	57.66	12.34	7.10	7.69
	Fe	3.39	10.47	31.77	37.32	7.88	9.16
	Di	7.38	10.17	6.48	4.31	56.78	14.87
	An	4.21	15.49	10.71	3.33	19.71	46.54
		112.4	93.5	125.2	63.8	109.1	96.0
Ma	Ha	80.45	4.71	6.14	3.36	4.07	1.28
	Sa	10.26	41.24	12.36	13.74	14.04	8.36
	Su	9.79	9.89	39.50	31.27	6.75	2.80
	Fe	5.12	10.64	22.85	48.29	8.09	5.01
	Di	14.55	14.23	6.75	10.90	44.61	8.96
	An	11.35	22.22	10.62	11.61	19.20	25.00
		131.5	102.9	98.2	119.2	96.8	51.4
Fe	Ha	70.26	5.02	9.01	2.08	9.95	3.67
	Sa	2.31	43.38	12.95	7.97	13.51	19.88
	Su	6.42	6.77	56.38	16.61	5.90	7.92
	Fe	1.82	9.81	28.36	40.65	8.69	10.66
	Di	4.72	10.99	6.91	6.68	52.67	18.04
	An	2.36	15.26	9.33	6.08	18.48	48.49
		87.9	91.2	122.9	80.0	109.2	108.7

B. Gender Differences in FER: Regions of Influence for Classification

Table IX and X present the merged heatmaps for the LIME images which help us understand those image regions that impact the classification of the image into a class. It is important to remark that even if heatmaps are similar between expressions, the network could be observing different features in those zones (e.g., the presence of frown or not). To focus on the differences in the important zones between male and female, we used the heatmaps of FE-FE and MA-MA to calculate their subtraction. We obtained three images per expression showing the absolute difference, the difference between FE-FE and MA-MA, and the difference between MA-MA and FE-FE (see Table XI).

Observing the difference in the heatmaps, the disparity between female and male datasets for the sadness, fear and anger expressions is lower than happiness, disgust and surprise, which are the most different regarding gender.

In the happiness expression, the network trained both with males and females give importance to the lower face region (see Table IX), but in female training, this zone is more highlighted, and in male training, cheeks are also important.

In the case of disgust, the female dataset (FE-FE) focuses on the lower region face (the mouth and chin), whereas male datasets (MA-MA) accentuate the central zone comprised between the mouth and the eyes (up to the temples) (see Table X).

In the surprise expression, similar heatmaps are achieved with both male and female datasets (see Table IX), however, female datasets (FE-FE) highlight with more intensity the upper-middle face.

Heatmaps for surprise and fear are quite similar (see Table IX), this could be the reason for their misclassifications (see Fig. 3,

TABLE IX. Heatmaps of Merged LIME Explanations for Each Expression, Training and Testing Datasets

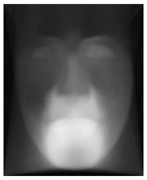
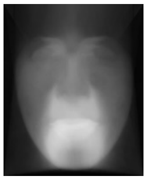
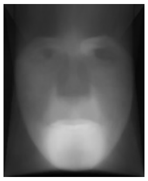
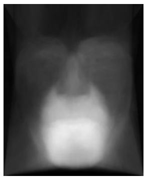
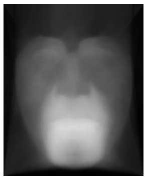
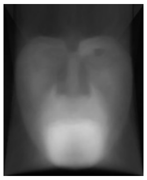
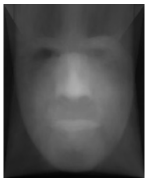
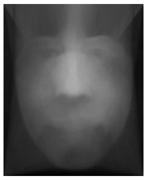
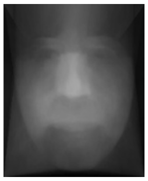
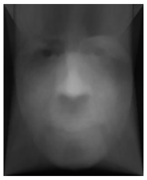
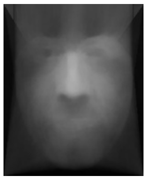
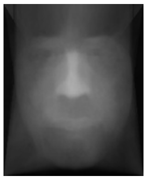
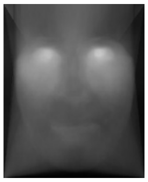
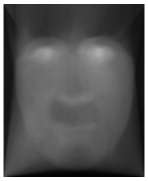
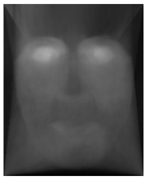
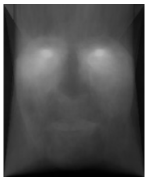
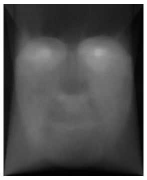
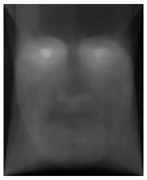
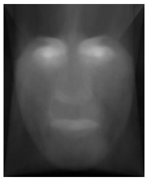
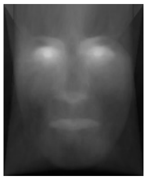
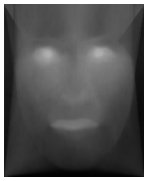
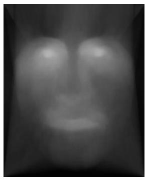
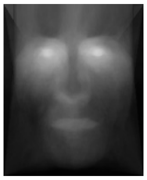
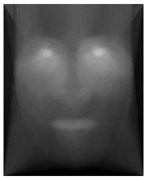
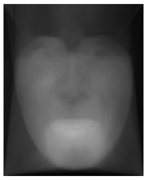
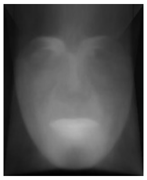
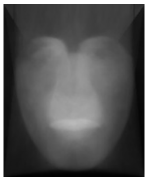
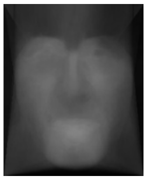
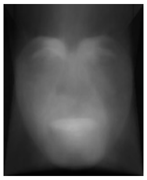
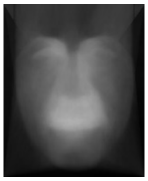
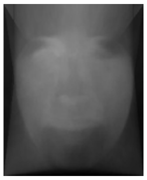
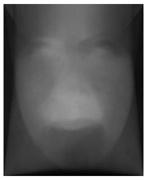
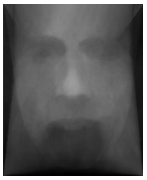
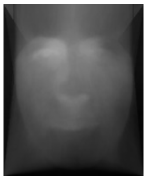
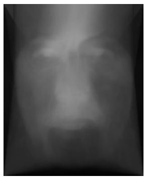
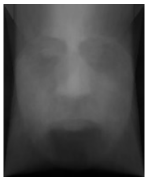
		Training Female	Training Mixed	Training Male
Happiness	Test Fe			
	Test Ma			
Sadness	Test Fe			
	Test Ma			
Surprise	Test Fe			
	Test Ma			
Fear	Test Fe			
	Test Ma			

TABLE X. Heatmaps of Merged LIME Explanations for Each Expression, Training and Testing Datasets (Cont.)

		Training Female	Training Mixed	Training Male
Disgust	Test Fe			
	Test Ma			
Anger	Test Fe			
	Test Ma			

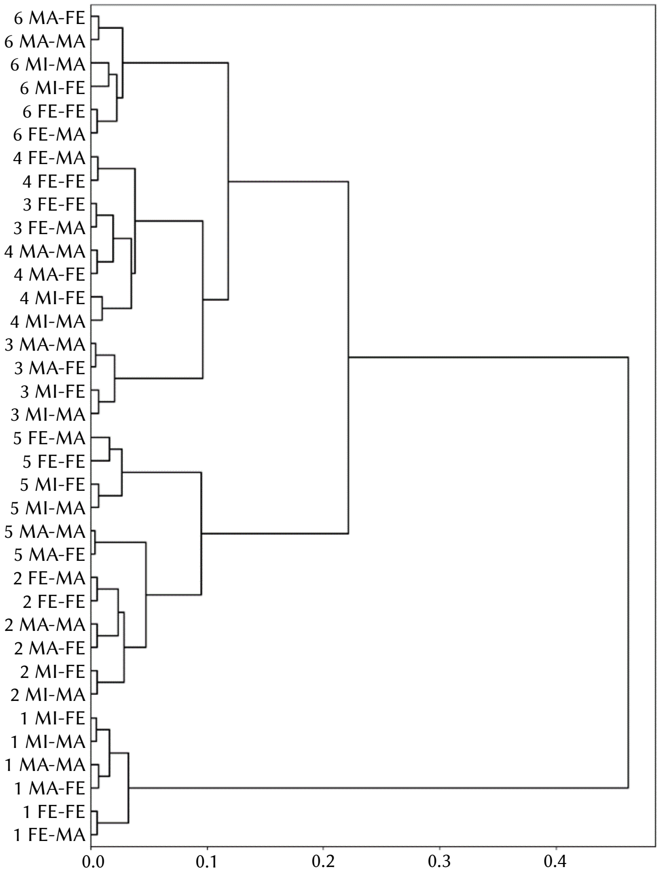
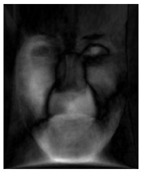
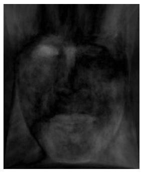
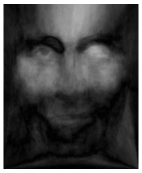
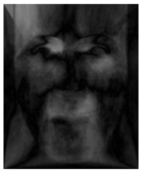
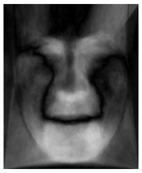
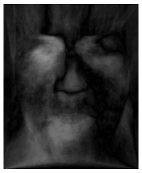
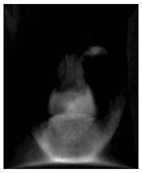
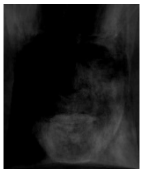
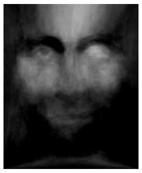
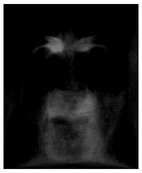
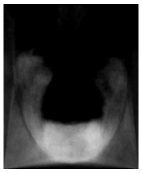
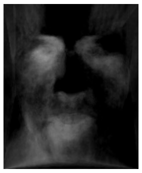
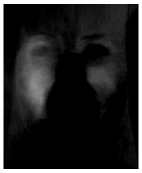
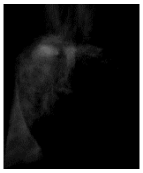
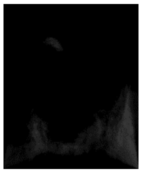
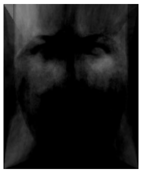
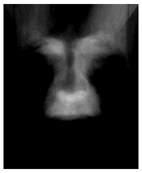
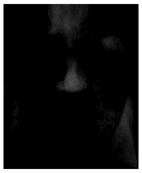


Fig. 3. Dendrogram with expressions happiness (1), sadness (2), surprise (3), fear (4), disgust (5) and anger (6). Ward's method is used to join clusters. Distance between clusters is computed as one minus normalized correlation.

TABLE XI. Differences Between Heatmaps for Each Expression. First Row: Absolute Value of the Difference Between the Heatmaps MA-MA and FE-FE. Second Row: Difference Between the Heatmaps FE-FE and MA-MA. Third Row: Difference Between the Heatmaps MA-MA and FE-FE. The Scalar Is Just Used to Improve the Visualization of the Differences. FE-FE Is Represented With F, MA-MA Is Represented With M

	Happiness	Sadness	Surprise	Fear	Disgust	Anger
3*|f-m|						
3*Max(f-m, 0)						
3*Max(m-f, 0)						

where the heatmaps for these expressions are grouped hierarchically based on similarity). Further, observing the grouped clusters in the dendrogram, happiness is the most different expression, which means that the important facial regions for learning are different to the other expressions and may help in the recognition.

The dendrogram also highlights the similarities of the heatmaps built using the mixed, male or female datasets for each expression. In general, expressions created with all datasets focus on similar regions, except for disgust trained with male datasets and surprise trained with female datasets. The first union of branches is mainly between those heatmaps belonging to the same training dataset and expression, meaning that the influence of the LIME images provided by the testing datasets is scarce. The next unions of branches depend mainly on the expression (with some exceptions, disgust in male training and surprise in female training), meaning that independently of the datasets, the expression influences the heatmap created. The similarities between heatmaps indicate that fear and surprise are the most similar and they share regions with anger, then sadness and disgust also coincide in face regions. And as already commented, happiness is the most different expression regarding the important face regions.

Although in some expressions an influence of the face regions indicated by Ekman [9] are shared (e.g. lower face for happiness), in general there is no direct relationship with them or even with AUs.

However, as Disgust is the most different expression regarding face regions, when planning a new study with limited resources, endeavors to achieve extra Disgust images from both genders can benefit FER. Summarizing, as highlighted regions are different in several expressions, the lack of diversity in the training dataset may impact the misclassification of facial expressions.

IV. Conclusion

Currently FER is a relevant area of research due to its wide range of applications, and frequently new approaches are validated using well-known datasets or models. However, especially in web-scrapped datasets, we cannot assure that they include an evenly distributed number of images of individuals in terms of sensitive attributes such as gender. Further, analyzing gender differences will help improve FER systems and understand the source of any inequality o misclassification. Considering this, we questioned: What are the gender differences when training a ML system for FER?

The aim of the present study was to study how gender impacts in the learning of a CNN. We explained comprehensively with XAI techniques the differences and similarities of the important face regions for the model to classify an expression into a class, considering gender.

A first contribution of the work is the novel explanation technique used for the comparison, that is, the creation of a unique heatmap based on the individual results of applying LIME on each image. By merging the instance-level information in a normalized image, we acquire global knowledge about the functioning of the CNN. Then, heatmaps can be used to calculate and analyze the similarities and differences among expressions and gender. This technique can be transferred to other similar FER studies.

Regarding FER, as expected, gender-balanced training datasets improve FER accuracy, achieving similar likelihood for positive and negative outcomes. However, unbalanced datasets do not affect in the same manner all the face expressions. Results show that training with male datasets achieve better results than training with female datasets, this could be related with the claims of women being generally more expressive than men and being better senders of nonverbal information [3], [54], therefore, the training with less exaggerated expressions could transfer better. However, the expression of anger is an exception, which can relate with the literature than men pose anger more intensely.

Analyzing the performance, the expression of happiness is globally well recognized, with the exception of female training tested on males. In addition, it is interesting how female training tends to classify frequently male images into the anger class, whereas, in the opposite way, male training tends to classify expressions by women as happiness. And the mixed training, decides frequently to classify images into other classes before using the fear expression.

Lastly, the findings of the comprehensive study of the important face regions for the neural network show that there exist common regions in some expressions both for females and males with different intensities (e.g., happiness); however, in expressions like disgust, face regions are different. Therefore, when datasets are not balanced, these differences can impact the correct classification of facial expressions. In addition, regarding the differences between gender and expression in the face regions important for the model, we observed that the expression is more influential than the training or testing datasets.

As Xu et al. [21] commented, there is a need for the research community to invest effort in creating facial expression datasets with explicit labels regarding sensitive attributes. The gender labelled file used in this study is available in https://github.com/josebambu/AffectNetGenderLabelling/.

We note that the results obtained are achieved with a particular relevant dataset (AffectNet dataset) and model (Inception), therefore, future work lines are to study if different datasets and neural networks behave similarly as this study and analyze if important facial regions for the network coincide with human perception, in order to build more human-based models. Further, it would be of interest to apply the methodology to other types of images (e.g. thermal [55] or depth [56]), and other fields of study such as face identification [57] or pain detection [27].

Acknowledgment

This work has been supported by the Agencia Estatal de Investigación, project PID2019-104829RA-I00 / MCIN/ AEI / 10.13039/501100011033, EXPLainable Artificial INtelligence systems for health and well-beING (EXPLAINING).

References

[1] D. McDuff, E. Kodra, R. el Kaliouby, and M. LaFrance, "A large-scale analysis of sex differences in facial expressions," *PLOS ONE*, vol. 12, no. 4, pp. 1–11, 2017.

[2] L. Cattaneo, V. Veroni, S. Boria, G. Tassinari, and L. Turella, "Sex Differences in Affective Facial Reactions Are Present in Childhood," *Frontiers in Integrative Neuroscience*, vol. 12, 2018.

[3] U. Dimberg and L.-O. Lundquist, "Gender differences in facial reactions to facial expressions.," *Biological Psychology*, vol. 30, no. 2. Elsevier Science, Netherlands, pp. 151–159, 1990.

[4] M. A. Kraines, L. J. A. Kelberer, and T. T. Wells, "Sex differences in attention to disgust facial expressions," *Cognition and Emotion*, vol. 31, no. 8, pp. 1692–1697, 2017.

[5] K. AM and G. AH., "Sex differences in emotion: expression, experience, and physiology," *J Pers Soc Psychol.*, vol. 74, no. 3, pp. 686–703, 1998.

[6] M. Thunberg and U. Dimberg, "Gender Differences in Facial Reactions to Fear-Relevant Stimuli," *Journal of Nonverbal Behavior*, vol. 24, no. 1, pp. 45–51, 2000.

[7] G. E. Schwartz, S. -L Brown, and G. L. Ahern, "Facial Muscle Patterning and Subjective Experience During Affective Imagery: Sex Differences.," *Psychophysiology*, vol. 17, pp. 75–82, 1980.

[8] C. Evers, A. H. Fischer, and A. S. R. Manstead, "Gender and emotion

regulation: a social appraisal perspective on anger.," in *Emotion regulation and well-being.*, Evers, Catharine: Department of Clinical and Health Psychology, Utrecht University, P.O. Box 80140, Utrecht, Netherlands, 3508 TC, c.evers@uu.nl: Springer Science + Business Media, 2011, pp. 211–222.

[9] P. Ekman, "Universals and cultural differences in facial expressions of emotion," *Nebraska Symposium on Motivation*, vol. 19, pp. 207–283, 1971.

[10] P. Ekman and W. . Friesen, *Facial action coding system: manual.* Palo Alto, Calif.: Consulting Psychologists Press. OCLC: 5851545, 1978.

[11] Y. Fan, J. C. K. Lam, and V. O. K. Li, "Demographic effects on facial emotion expression: an interdisciplinary investigation of the facial action units of happiness," *Scientific Reports*, vol. 11, no. 1, p. 5214, 2021.

[12] O. Houstis and S. Kiliaridis, "Gender and age differences in facial expressions," *European Journal of Orthodontics*, vol. 31, no. 5, pp. 459–466, 2009.

[13] T. S. H. Wingenbach, C. Ashwin, and M. Brosnan, "Sex differences in facial emotion recognition across varying expression intensity levels from videos," *PloS one*, vol. 13, no. 1, pp. e0190634–e0190634, Jan. 2018.

[14] B. Montagne, R. P. C. Kessels, E. Frigerio, E. H. F. de Haan, and D. I. Perrett, "Sex differences in the perception of affective facial expressions: Do men really lack emotional sensitivity?," *Cognitive Processing*, vol. 6, no. 2, pp. 136–141, 2005.

[15] R. Campbell *et al.*, "The classification of 'fear' from faces is associated with face recognition skill in women," *Neuropsychologia*, vol. 40, no. 6, pp. 575–584, 2002.

[16] A. K. Vail, J. F. Grafsgaard, K. E. Boyer, E. N. Wiebe, and J. C. Lester, "Gender Differences in Facial Expressions of Affect During Learning," in *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, 2016, pp. 65–73.

[17] J. C. Borod, E. Koff, and B. White, "Facial asymmetry in posed and spontaneous expressions of emotion," *Brain and Cognition*, vol. 2, no. 2, pp. 165–175, 1983.

[18] W. Mellouk and W. Handouzi, "Facial emotion recognition using deep learning: review and insights," *Procedia Computer Science*, vol. 175, pp. 689–694, 2020.

[19] A. Domnich and G. Anbarjafari, "Responsible AI: Gender bias assessment in emotion recognition," *arXiv*, pp. 1–19, 2021.

[20] M. Deramgozin, S. Jovanovic, H. Rabah, and N. Ramzan, *A Hybrid Explainable AI Framework Applied to Global and Local Facial Expression Recognition.* 2021.

[21] T. Xu, J. White, S. Kalkan, and H. Gunes, "Investigating Bias and Fairness in Facial Expression Recognition," in *ECCV Workshops 2020*, 2020.

[22] Z. Wang *et al.*, "Towards Fairness in Visual Recognition: Effective Strategies for Bias Mitigation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 8916-8925.*, 2020.

[23] M. Lyons, M. Kamachi, and J. Gyoba, "The Japanese Female Facial Expression (JAFFE) Dataset." Zenodo, 1998.

[24] A. Heimerl, K. Weitz, T. Baur, and E. Andre, "Unraveling ML Models of Emotion with NOVA: Multi-Level Explainable AI for Non-Experts," *IEEE Transactions on Affective Computing*, vol. 1, no. 1, pp. 1–13, 2020.

[25] D. Schiller, T. Huber, M. Dietz, and E. André, "Relevance-Based Data Masking: A Model-Agnostic Transfer Learning Approach for Facial Expression Recognition," *Frontiers in Computer Science*, vol. 2, p. 6, 2020.

[26] P. Prajod, D. Schiller, T. Huber, and E. Andr'e, "Do Deep Neural Networks Forget Facial Action Units? - Exploring the Effects of Transfer Learning in Health Related Facial Expression Recognition," *ArXiv*, vol. abs/2104.0, 2021.

[27] K. Weitz, T. Hassan, U. Schmid, and J.-U. Garbas, "Deep-learned faces of pain and emotions: Elucidating the differences of facial expressions with the help of explainable AI methods," *Technisches Messen*, vol. 86, no. 7–8, pp. 404–412, 2019.

[28] G. del Castillo Torres, M. F. Roig-Maimó, M. Mascaró-Oliver, E. Amengual-Alcover, and R. Mas-Sansó, "Understanding How CNNs Recognize Facial Expressions: A Case Study with LIME and CEM," *Sensors*, vol. 23, no. 1, 2023.

[29] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[30] M. Alber *et al.*, "iNNvestigate Neural Networks!," *Journal of Machine Learning Research*, vol. 20, no. 93, pp. 1–8, 2019.

[31] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation," *PLOS ONE*, vol. 10, no. 7, pp. 1–46, 2015.

[32] C. Manresa-Yee and S. Ramis, "Assessing Gender Bias in Predictive Algorithms Using EXplainable AI," in *Proceedings of the XXI International Conference on Human Computer Interaction*, 2021.

[33] K. Grabowski *et al.*, "Emotional expression in psychiatric conditions: New technology for clinicians," *Psychiatry and Clinical Neurosciences*, vol. 73, no. 2, pp. 50–62, 2019.

[34] A. M. Barreto, "Application of facial expression studies on the field of marketing," *Emotional expression: the brain and the face*, vol. 9, no. June, pp. 163–189, 2017.

[35] S. Medjden, N. Ahmed, and M. Lataifeh, "Adaptive user interface design and analysis using emotion recognition through facial expressions and body posture from an RGB-D sensor," *PLoS ONE*, vol. 15, no. 7, p. e0235908, 2020.

[36] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 01, pp. 18–31, 2019.

[37] P. Ekman, "An argument for basic emotions," *Cognition and Emotion*, vol. 6, no. 3–4, pp. 169–200, 1992.

[38] S. Li and W. Deng, "Reliable Crowdsourcing and Deep Locality-Preserving Learning for Unconstrained Facial Expression Recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 356–370, 2019.

[39] World Health Organization, "Gender," *Available online: https://www.who.int/europe/health-topics/gender.* .

[40] Y. Chen and J. Joo, "Understanding and Mitigating Annotation Bias in Facial Expression Recognition," in *ICCV 2021*, 2021.

[41] J.-L. Lisani, S. Ramis, and F. Perales, "A Contrario Detection of Faces: A Case Example," *SIAM Journal on Imaging Sciences*, vol. 10, pp. 2091–2118, Jan. 2017.

[42] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 Faces in-the-Wild Challenge: The First Facial Landmark Localization Challenge," in *2013 IEEE International Conference on Computer Vision Workshops*, 2013, pp. 397–403.

[43] S. Ramis, J. . Buades, F. J. Perales, and C. Manresa-Yee, "A Novel Approach to Cross dataset studies in Facial Expression Recognition," *Multimedia Tools and Applications.*

[44] A. Barredo Arrieta *et al.*, "Explainable Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, no. October 2019, pp. 82–115, 2020.

[45] J. Ward, "Hierarchical Grouping to Optimize an Objective Function," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236–244, 1963.

[46] O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[47] C. Szegedy *et al.*, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9.

[48] M. Lin, C. Qiang, and Y. Shuicheng, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[49] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning with Applications in R.* Springer, New York, 2013.

[50] D. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*, 2015.

[51] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, "Suppressing uncertainties for large-scale facial expression recognition," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 6896–6905, 2020.

[52] Q. T. Ngo and S. Yoon, "Facial expression recognition based on weighted-cluster loss and deep transfer learning using a highly imbalanced dataset," *Sensors (Switzerland)*, vol. 20, no. 9, 2020.

[53] C.-T. Yen and K.-H. Li, "Discussions of Different Deep Transfer Learning Models for Emotion Recognitions," *IEEE Access*, vol. 10, pp. 102860–102875, 2022.

[54] H. G. Wallbott, “Big girls don’t frown, big boys don’t cry—Gender differences of professional actors in communicating emotion via facial expression.,” *Journal of Nonverbal Behavior*, vol. 12, no. 2, pp. 98–106, 1988.

[55] N. K. Benamara, E. Zigh, T. B. Stambouli, and M. Keche, “Towards a Robust Thermal-Visible Heterogeneous Face Recognition Approach Based on a Cycle Generative Adversarial Network,” *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 4, pp. 132–145, 2022.

[56] M. Z. Uddin, M. M. Hassan, A. Almogren, M. Zuair, G. Fortino, and J. Torresen, “A facial expression recognition system using robust face features from depth videos and deep learning,” *Computers & Electrical Engineering*, vol. 63, pp. 114–125, 2017.

[57] A. Alcaide, M. A. Patricio, A. Berlanga, A. Arroyo, and J. J. Cuadrado-Gallego, “LIPSNN: A Light Intrusion-Proving Siamese Neural Network Model for Facial Verification,” *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 4, pp. 121–131, 2022.

Cristina Manresa-Yee

Cristina Manresa-Yee received her degree in Computer Science and her Ph. D. in Computer Science from the University of Balearic Islands. She is currently an Associate Professor at the University of the Balearic Islands. Her research interests include human-computer interaction, computer vision and explainable AI.

Silvia Ramis

Silvia Ramis, Ph. D. in Information and Communications Technologies from the UIB (since 2019). She has participated in several projects in the field of Computer Vision, Artificial Intelligence, Explainable Artificial Intelligence and Human-Robot Interaction. Her research experience focuses on artificial intelligence applied to human-robot interaction, especially in face detection and facial expression recognition.

Jose M. Buades

Jose Maria Buades Rubio received his degree in Computer Science and his Ph. D. in Computer Science from the University of Balearic Islands. He is currently an Associate Professor at the University of the Balearic Islands. His research interests include computer graphics, computer vision and artificial intelligence.

A Feature Selection Approach Based on Archimedes' Optimization Algorithm for Optimal Data Classification

Lahbib Khrissi[1]*, Nabil El Akkad[2], Hassan Satori[1], Khalid Satori[1]

[1] LISAC, Department of Computer Science, Faculty of Science, Dhar El Mahraz, USMBA, Fez (Morocco)
[2] Laboratory of Engineering, Systems and Applications, ENSA of Fez, USMBA, Fez (Morocco)

* Corresponding author: lahbib.khrissi@usmba.ac.ma

Received 25 February 2022 | Accepted 2 December 2022 | Published 19 January 2023

Abstract

Feature selection is an active research area in data mining and machine learning, especially with the increase in the amount of numerical data. FS is a search strategy to find the best subset of features among a large number of subsets of features. Thus, FS is applied in most modern applications and in various domains, which requires the search for a powerful FS technique to process and classify high-dimensional data. In this paper, we propose a new technique for dimension reduction in feature selection. This approach is based on a recent metaheuristic called Archimedes' Optimization Algorithm (AOA) to select an optimal subset of features to improve the classification accuracy. The idea of the AOA is based on the steps of Archimedes' principle in physics. It explains the behavior of the force exerted when an object is partially or fully immersed in a fluid. AOA optimization maintains a balance between exploration and exploitation, keeping a population of solutions and studying a large area to find the best overall solution. In this study, AOA is exploited as a search technique to find an optimal feature subset that reduces the number of features to maximize classification accuracy. The K-nearest neighbor (K-NN) classifier was used to evaluate the classification performance of selected feature subsets. To demonstrate the superiority of the proposed method, 16 benchmark datasets from the UCI repository are used and also compared by well-known and recently introduced meta-heuristics in this context, such as: sine-cosine algorithm (SCA), whale optimization algorithm (WOA), butterfly optimization algorithm (BAO), and butterfly flame optimization algorithm (MFO). The results prove the effectiveness of the proposed algorithm over the other algorithms based on several performance measures used in this paper.

Keywords

Archimedes' Optimization Algorithm, Data Classification Feature Selection, KNN Classifier.

DOI: 10.9781/ijimai.2023.01.005

I. Introduction

Over recent years, storage capacity and information acquisition have become cheaper. This allows us to store all possible data related to people's needs. However, there are data that do not necessarily contain useful information to be extracted. With the exceptional increase in the amount of data that can be stored, the exploration of new methods that are able to process data automatically is necessary following a process of knowledge extraction from data. This process allows for the integration and collection of data, selection, cleaning and data processing, data analysis for the extraction of appropriate patterns and models, evaluation and interpretation of constructed models and consolidation of knowledge available for use.

Feature selection (FS) is a research process or technique used to select the most interesting, relevant, or informative features, variables, or measures of a given system in order to accomplish the task for which it was designed [1],[2]. In the field of machine learning and more particularly in classification, some irrelevant and/or redundant features, usually present in the training data, not only make learning more difficult, but also degrade the generalization performance of the training models [3],[4]. FS is a pre-processing step that plays an important role in data mining. It allows to represent a subset of data from a large data set and to eliminate redundant, irrelevant or noisy data. There are several advantages of attribute subset selection: It facilitates data visualization and provides a better understanding [5],[6]. It reduces the complexity of the training data which will lead to the reduction of the time of the learning algorithm. Another important factor is the reduction of the problem dimension, the improvement of the prediction performance and the understanding of the learning model [7],[8]. FS methods are applied to several applications and in various well-known fields, such as computer vision [9],[10], image processing [11]–[13] pattern recognition and machine learning [14]–[16].

Classification algorithms aim to identify the classes to which objects belong based on certain descriptive features. So, the main process of classification in machine learning is to train the classifier to accurately recognize patterns from given training samples and classify the test

Please cite this article as:
L. Khrissi, N. E. Akkad, H. Satori, K. Satori. A Feature Selection Approach Based on Archimedes' Optimization Algorithm for Optimal Data Classification, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 1, pp. 28-38, 2024, http://dx.doi.org/10.9781/ijimai.2023.01.005

samples with the trained classifier. The choice of a classification method for the treatment of the FS problem by metaheuristics has a great impact on the quality of the obtained solution. In the literature, there are several types of classifiers that have been used to assist data selection techniques [17] among those commonly used we mention: support vector machine (SVM), K-nearest neighbor (KNN), naive bay (NB), random forest (RF), artificial neural network (ANN) and others.

Metaheuristics (MH) include all algorithms based on the population concept that use selection and recombination to generate new points in the search space. MHs are widely used to solve complex optimization problems. They have been developed and updated for use in various domains, for example task scheduling in cloud computing [18], bioinformatics [19], feature selection [20], image segmentation [21]–[23] and camera self-calibration [24],[25]. However, all MHs need to properly balance the exploration and exploitation phases to achieve good results, otherwise the solutions tend to be trapped in local optima or cannot converge properly.

MHs have been widely adopted to solve complex optimization tasks, including FS. Thus, the FS process for classification can be viewed as a search problem in a state space, where each state can be represented as a vector of size equal to the number of attributes in the problem and each element of the vector can take the value 1 if the corresponding attribute is selected and 0 otherwise. In the literature several metaheuristics have been widely used in recent years to deal with the FS optimization process. Such as, ant optimization algorithm [26], dragonfly optimization algorithm [27], whale optimization algorithm (WOA) [28] marine predators' algorithm (MPA)[29], sine cosine algorithm (SCA) [30], [31] Harris hawk optimization (HHO) algorithm[32] and other. Thus, there are also hybridizations between metaheuristics that has been used to address the FS problem [33]–[35].

Lately (in 2021), a new MH proposed by Hashim et al [36] to solve real-world problems called the Archimedes' optimization algorithm (AOA). The latter is based on Archimedes' principle in the law of physics. It describes the behavior of the force exerted when an object is partially or totally immersed in a fluid. Like most MHs, the AOA proposes solutions in the form of a population. Thus, the search agents are the immersed objects. Their density, volume and acceleration during collisions with other nearby objects are updated to define the updated positions of the objects. The main objective of the AOA is to bring the objects to a state of equilibrium. AOA was tested on 29 reference functions and with four engineering design problems. The results obtained by AOA showed promising results, and shows that AOA balanced between the phases of exploration and exploitation. The search performance of this algorithm was compared to well-known algorithms like GA, PSO, as well as to more recent additions such as WOA, SCA, HHO and EO.

AOA has already started to attract the attention of several researchers for the use of this algorithm in a variety of optimization problems, thanks to their important criteria related to simplicity, efficiency, adaptability and flexibility. In the literature, several works that are used AOA as an optimizer, among them we can cite: in [37] for human facial analysis, in [38] to solve the optimal locations and sizes of solar photovoltaic systems (SPV) in electrical distribution networks, in [39] for wind speed prediction, in [40] to eliminate selective harmonics in a cascaded H-bridge inverter (CHB). In addition, researchers have introduced modifications to the original AOA to increase its effectiveness. In the literature there are several improved versions of the traditional AOA for example: Enhanced Archimedes Optimization Algorithm (EAOA)[41] to improve the balance between exploration and exploitation of AOA and to improve the classification performance, I-AOA [42] an improved version of AOA which is based on combination of two effective strategies, the local escape operator (LEO) and orthogonal learning (OL) to determine the optimal parameters of the polymer electrolyte membrane (PEM) fuel cell (FC) and IAOA [43] allows to increase the population diversity in AOA, to further improve the balance between exploitation and exploration of AOA, and to avoid premature convergence problems whose objective to solve the optimal power flow problem (OPF).

According to the No-Free-Lunch (NFL) theorem, there is no algorithm that solves all optimization problems [44], i.e., most metaheuristics fail when the problem is modified. Moreover, the results obtained by AOA are promising and the statistical analysis of these results revealed that this algorithm can be considered as a good performance optimizer compared to well-known optimization algorithms such as: Multi-verse Optimizer (MVO) [45], Henry Gas Solubility Optimization (HGSO)[20], Harris Hawks Optimize (HHO) [46], Equilibrium Optimizer (EO) [47], and others. Moreover, AOA is used as an optimizer in several works mentioned above, which shows its optimization efficiency that is ensured by many important criteria that are related to simplicity, efficiency, adaptability and flexibility. This motivated us to try to develop a new FS approach based on AOA. The treatment of the FS problem by this algorithm aims at finding a good classification with high accuracy.

In this work, a new proposed FS approach based on envelopes by AOA is intended for data classification. It should be noted that this algorithm maintains a perfect balance between exploitation and exploration. Due to this feature, AOA is well suited for solving complex problems, especially feature selection. Our method is based on a hybridization between the AOA algorithm and k-NN for feature selection and data classification; the system we propose is defined in three phases: The first is the initialization phase, we generate a number of objects as the initial population and the size of each object in AOA corresponds to the number of attributes; the second is the phase of solution updates, we evaluate the quality of each candidate solution using AOA-assisted enveloping feature selection, a good compromise between accuracy and a reduced number of features should be ensured by a proposed objective function; and the last one is the classification phase, after finishing the process of our method, we return the best solution for classification, in this paper we used the K-NN classifier.

The important contributions of this paper are:

- The introduction of a new algorithm for the FS problem based on AOA.
- The performance of AOA was compared to other well-known metaheuristic algorithms in the literature for FS, such as MFO, SCA, WOA and BOA, all based on accuracy and number of selected features.
- The proposed method is evaluated on sixteen datasets with significantly high dimensions and small instances.
- The classification of the data by the features selected by the different optimizers was performed by the KNN classifier.
- A statistical comparison of the obtained results was performed by the optimizers using different analysis metrics.

The remaining sections of this paper are organized as follows: Section II will present the overview on the algorithms used in this work. The proposed method is described in Section III to solve the classification problem. The experiments and the results obtained are discussed in section V. Finally, the conclusion is presented in section VI.

II. Background Overview

A. Archimedes' Optimization Algorithm (AOA)

The AOA algorithm is a recent population-based metaheuristic introduced by Fatma Hashim in 2020 [36] This algorithm is part of the metaheuristics that are inspired by the rules of physics of the universe, especially Archimedes' law. The principle of Archimedes' phenomenon states that when an object is completely or partially immersed in a fluid, the fluid exerts an upward force on the object equal to the weight of the fluid displaced by the object [48]. The importance of the AOA algorithm lies in the formulation of the solution which is based on three auditory information: Volume (V), Density (D) and Acceleration (A) to the base agents. Thus, initially, the group of agents is randomly generated with dimensions (Dim). The random values V, D and A are provided as they are additive data. Then, the evaluation mechanism is performed for each object to specify the best object (O_{best}).

The operation of this algorithm is realized as follows: At each iteration, the AOA updates the density and volume of each object. The acceleration of the object is updated according to the condition of its collision with any other nearby object. The updated density, volume and acceleration determine the new position of an object in the current solution. The main steps of the AOA are described below:

1. **Initialization:** This step randomly initializes the real population that contains N objects using equation (1). Thus, each object is characterized by its density (D_i), volume (V_i) and acceleration (Γ_i) which are randomly defined using the following equations: Eq. (2), Eq. (3) and Eq. (4):

$$O_i = O_i^{min} + r_1 \times \left(O_i^{max} - O_i^{min}\right); \;\; i = 1, 2, \ldots., N \qquad (1)$$

$$D_i = r_2 \qquad (2)$$

$$V_i = r_3 \qquad (3)$$

$$\Gamma_i = \Gamma_i^{max} + r_4 \times \left(\Gamma_i^{max} - \Gamma_i^{min}\right); \;\; i = 1, 2, \ldots., N \qquad (4)$$

Where O_i represents the ith object, O_i^{max} and O_i^{min} are the maximum and minimum boundaries of the search space, respectively.

r_1, r_2, r_3 and r_4 are random vectors that belong to [0, 1].

The population will be evaluated by calculating the score of each object to unearth the best object (O_{best}) by combining their best values of density (D_{best}), volume (V_{best}) and acceleration (Γ_{best}).

2. **Density and volume updates:** In this step, the density and volume values for each object are updated by checking the best density and volume using Eq. (5) and Eq. (6).

$$D_i^{t+1} = D_i^t + s_1 \times (D_{Best} - D_i^t) \qquad (5)$$

$$V_i^{t+1} = V_i^t + s_2 \times (V_{Best} - V_i^t) \qquad (6)$$

Where s_1, s_2 are random scalars in [0, 1].

3. **Transfer coefficient and density scalar:**

In this step, the collision between the objects occurs until the steady state is reached. The main role of the transfer function ( ) is to switch from exploration mode to exploitation mode, defined by equation (7):

$$T_c = \exp\left(\frac{t-T}{T}\right) \qquad (7)$$

T_c increases exponentially with time until it reaches 1. t is the current iteration, while T denotes the maximum number of iterations. Moreover, decreasing the density scalar d_s in AOA allows us to find an optimal solution using Eq. (8):

$$d_s^{t+1} = exp\left(\frac{t-T}{T}\right) - \left(\frac{t}{T}\right) \qquad (8)$$

4. **Exploration phase:** In this step, the collision between agents occurs using random material selection (Mr). Thus, the update of acceleration objects is applied using equation (9) when the value of the transfer function is less than or equal to 0.5.

$$\Gamma_i^{t+1} = \frac{D_{Mr} + V_{Mr} + \Gamma_{Mr}}{D_i^{t+1} + V_i^{t+1}} \qquad (9)$$

5. **Exploitation phase:** In this step, the collision between the agents is not realized. Thus, the update of acceleration objects is applied using equation (10) when the value of the transfer coefficient is greater than 0.5.

$$\Gamma_i^{t+1} = \frac{D_{Best} + V_{Best} + \Gamma_{Best}}{D_i^{t+1} + V_i^{t+1}} \qquad (10)$$

Where Γ_{best} is the acceleration of the optimal object O_{best}.

6. **Acceleration normalization:** In this step, we normalize the acceleration to determine the rate of change using (11):

$$\Gamma_{i-norm}^{t+1} = \alpha \times \frac{\Gamma_i^{t+1} - \Gamma^{min}}{\Gamma^{max} - \Gamma^{min}} + \beta \qquad (11)$$

Where α and β are set to 0.9 and 0.1, respectively. The Γ_{i-norm}^{t+1} determines the percentage of steps each agent will change. The highest value of the acceleration means that the object performs the exploration operation; otherwise, the exploitation mode is operational.

7. **The update process:** For the exploration phase ($T_c \leq 0.5$), the position of the ith object at iteration t+1 is changed by equation (12), while the position of the object is updated by equation (13) in the exploitation phase ($T_c > 0.5$).

$$O_i^{t+1} = O_i^t + c_1 \times r_5 \times \Gamma_{i-norm}^{t+1} \times d_s \times (O_{rand} - O_i^t) \qquad (12)$$

Where c_1 is equal to 2.

$$O_i^{t+1} = O_{Best}^t + F \times c_2 \times r_6 \times \Gamma_{i-norm}^{t+1} \times d_s \times (\delta \times O_{Best} - O_i^t) \qquad (13)$$

Where c_2 is fixed at 6.

The parameter δ is positively correlated with time and this parameter is proportionally related to the transfer coefficient T_c, i.e. $\delta = 2 \times T_c$. The main role of this parameter is to ensure a good balance between exploration and exploitation operations. During the early iterations, the margin between the best object and the other object is higher, allowing for a high random walk. However, in the later iterations, the margin will be reduced and provide a low random walk.

F is used for marking which controls the search direction using equation (14):

$$F = \begin{cases} +1 & if \;\; \lambda \leq 0.5 \\ -1 & if \;\; \lambda > 0.5 \end{cases} \qquad (14)$$

Where λ = 2 × rand - 0.5.

8. **Evaluation:** In this step, we evaluate the new population using the score index Sc to determine the best object O_{best} and the best additive information, including D_{best}, V_{best} and Γ_{best}.

The pseudo code steps of the AOA algorithm is described in Algorithm 1.

B. K-Nearest Neighbor (K-NN)

The K-Nearest Neighbor (K-NN) is a method based on the notion of proximity (neighborhood) between variables and on reasoning from similar cases to make a decision. It is the training sample, associated with a distance function and a class choice function based on the classes of the nearest neighbors, which constitutes the model. To predict the class of a sample, the algorithm looks for the k nearest neighbors of this new case and predicts the most frequent answer of this k nearest neighbors [49]. Thus, the decision principle is simply to compute the distance of this sample to all the provided samples.

KNN is one of the most commonly used machine learning techniques with different datasets due to its simplicity and easy-to-implement advantages over other supervised machine learning

Algorithm 1: AOA

```
Initialization: N population size; T maximum iteration; $c_1$; $c_2$
Initialize N objects with their densities ($D$), volumes ($V$) and accelerations ($\Gamma$) using the equations from (1) to (4) respectively.
Evaluate the score for each object.
Determination of best object ($O_{Best}$) with ($D_{Best}$), ($V_{Best}$) and ($\Gamma_{Best}$)
Set t = 1
while t ≤ T do
    for each object i do
        Update density and volume using Eq.(5) and Eq.(6)
        Update transfer coefficient ($T_c$) and density scalar ($d_s$) using Eq. (7) and Eq. (8)
            if $T_c \leq 0.5$ then        (Exploration operation)
                Update acceleration ($\Gamma_i$) using Eq. (9)
                Normalize acceleration ($\Gamma_i$) using Eq. (11)
                Update position using Eq. (12)
            else                          (Exploitation operation)
                Update acceleration using Eq. (10)
                Update flagging control F using Eq. (13)
                Update position using Eq. (14)
            end if
            Compute the score of each object.
    end for
    Determine the best object ($O_{Best}$) with the best value of ($D_{Best}$), ($V_{Best}$) and ($\Gamma_{Best}$).
    Set $t = t + 1$
end while
return Best object with their quality
```

techniques. Therefore, KNN classifier is frequently used in several fields such as: healthcare, image and video recognition, finance, etc.

An object is classified according to a majority vote by its neighbors; the object obtains the class which is the most common among its K closest neighbors in the feature space. K must therefore be a positive integer, usually small. An odd k is often chosen to avoid equality in voting. The distance used for the calculation of the proximity of the neighbors is most often the Euclidean distance. The main steps of the k -NN algorithm for ranking the sample are presented in Algorithm 2.

Algorithm 2: k-NN algorithm

```
Inputs: Load the training and test data.
Outputs: Assign a class to the test point based on the majority of
    classes presented in the chosen points calculate accuracy
Choose the value of k for each point in test data
while stopping condition is not met do
    Find the Euclidean distance to all training data points.
    Store the Euclidean distances in a list and sort it.
    Choose the first k points.
Return Accuracy (Acc)
```

III. Proposed Method

In this section, the proposed approach has been discussed in detail. In addition, a flowchart and an algorithmic model have also been described to understand the proposed solution. Fig. 1 describes the general flowchart of the proposed method and its implication in the feature selection problem for real-world data sets. The AOA technique used focuses on finding an optimal subset of features from the training set and is tested using the validation set.

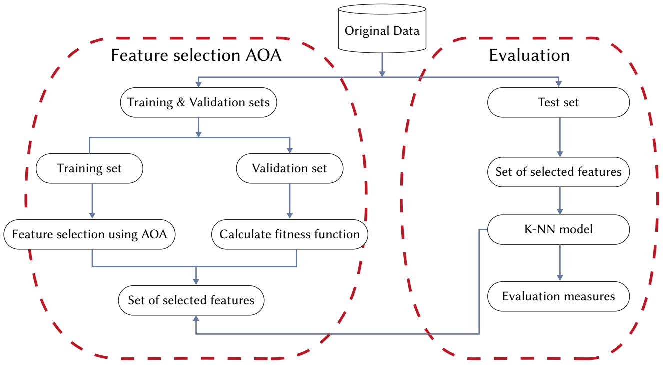


Fig. 1. The general flowchart of the proposed method.

According to the shortcomings of existing FS algorithms, this paper proposes a new method which is based on hybridization between AOA and k-NN algorithm for feature selection and data classification. The system we propose to solve the FS problem is defined in three phases: The first is the initialization phase, the second is the solution updates phase and the last is the classification phase.

In the first step, we define a solution as a numerical vector, we use a vector of (0 and 1) with 1 meaning that the attribute is selected, and 0 otherwise. We generate a number of objects as the initial population and the size of each object in the AOA is the number of attributes. At this point, if the value is greater than or equal to 0.5, then it is rounded to one. In this case, the attribute is considered a relevant feature. Conversely, the attribute is ignored when the value is rounded to zero.

In the second step, we evaluate the quality of each candidate solution using AOA-assisted wraparound feature selection, a good compromise between accuracy and a reduced number of features must be ensured by the objective function proposed in our paper which is described in equation (15) to determine the best $O_{best}O_{best}$ solution. Then, this process is repeated until the termination condition is met to make the necessary updates of the solutions.

In the third step, after finishing the process of our method, we

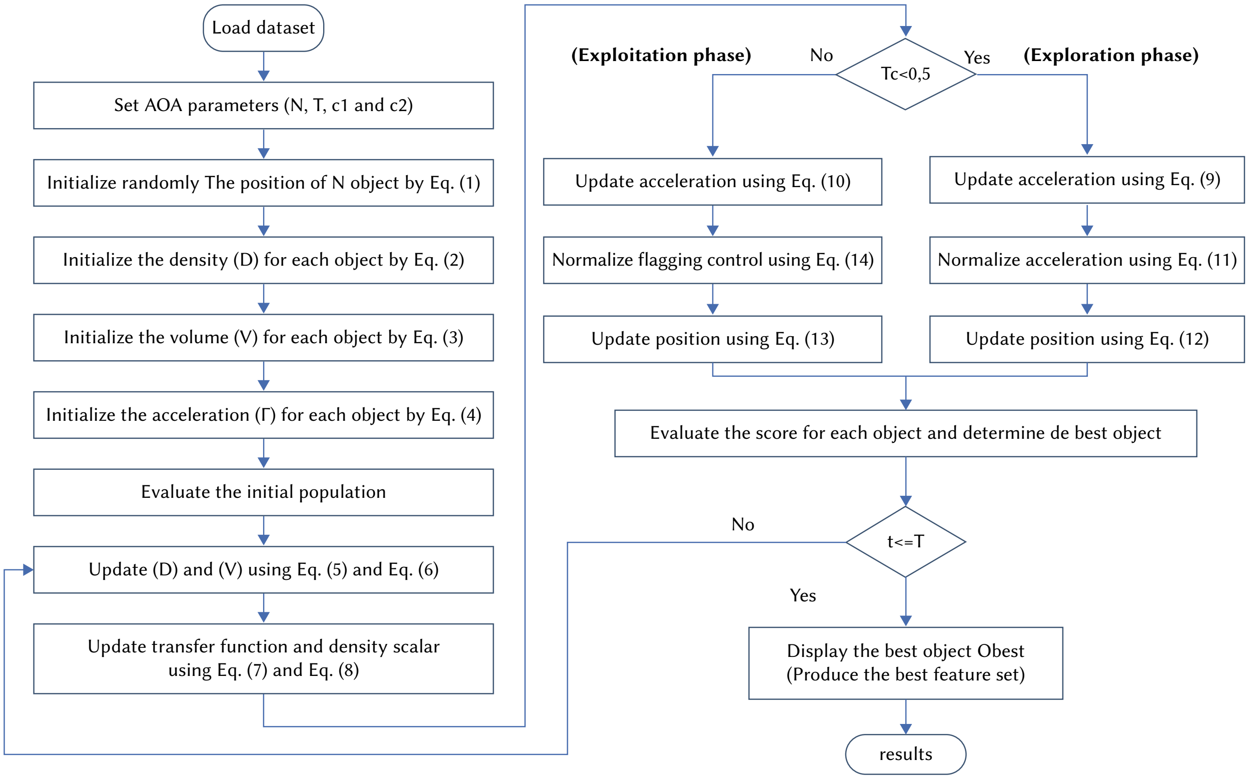

Load dataset
Set AOA parameters (N, T, c1 and c2)
Initialize randomly The position of N object by Eq. (1)
Initialize the density (D) for each object by Eq. (2)
Initialize the volume (V) for each object by Eq. (3)
Initialize the acceleration (Γ) for each object by Eq. (4)
Evaluate the initial population
Update (D) and (V) using Eq. (5) and Eq. (6)
Update transfer function and density scalar using Eq. (7) and Eq. (8)
(Exploitation phase)
No
Tc<0,5
Yes
(Exploration phase)
Update acceleration using Eq. (10)
Normalize flagging control using Eq. (14)
Update position using Eq. (13)
Update acceleration using Eq. (9)
Normalize acceleration using Eq. (11)
Update position using Eq. (12)
Evaluate the score for each object and determine de best object
No
t<=T
Yes
Display the best object Obest (Produce the best feature set)
results

Fig. 2. The flowchart of the proposed FS algorithm.

return the best solution O_{best}. In the original data, we keep only the features with their values corresponding to O_{best}. We used a retention strategy for classification, which implies that we randomly divide the data set into two parts: 80% for the training set and 20% for the test set. In this paper we choose the k -NN algorithm (k = 5) to evaluate the accuracy using a test set. The value of k in k -NN is set to k = 5 based on several papers in the literature to make a fair comparison [50],[51].

The steps for integrating the operators of the AOA algorithm into the feature selection operation are explained in the following flowchart (Fig. 2).

A. Fitness Function

The main objective of the FS problem is to maximize the classification performance and maintain a minimum number of selected features. In wrapper methods, the fitness function is related to the construction of a new classifier based on the features involved in the individual. And as we mentioned before, our method uses a wrapper, so a learning algorithm must be integrated in the evaluation process. In our study we used a classifier well known for its performances, it is the k-NN classifier. To circumvent the cumbersome nature of this approach, we have defined a fitness function that allows us to control the accuracy of the selected features during an iterative process to check the quality for each iteration. Therefore, the overall goal is to find the minimum value of the fitness function given in equation (15):

$$fit_{\vartheta} = \tau * Err + (1-\tau)\frac{\sum_i \vartheta_i}{n} \qquad (15)$$

Where:

- *Err* is defined as the classification error rate;
- τ is a constant controlling the classification importance with respect to the number of selected features;
- ϑ is a vector of size n with 0/1 elements representing unselected/selected features;
- *n* is the total number of features in the data set.

In general, wraparound methods search the space of subsets of variables, guided by the model output. They therefore incorporate the classification algorithm into the attribute selection procedure and use the classification error rate as an evaluation criterion. As mentioned earlier, our attribute selection approach is an envelope approach, which uses the k-NN classifier in the evaluation phase to correctly identify the appropriate features that should be selected. This method achieves good performance and often gets better results; however, it increases the time needed to reach a good solution. In our study the k-NN algorithm used in the trial-and-error experiments where the best choice of K is selected (K = 5) as the best performing on all data sets.

As shown in Fig. 3, our approach contains two search phases: local and global. The transition between these two phases is controlled by the transfer function given in equation (Eq. 7), which means that if the value of T_c is less than 0.5, our algorithm will perform a global search (exploration); otherwise, it will perform a local search (exploitation). Thus, our proposed method has achieved better efficiency when balancing between exploitation and exploration techniques in the search space.

IV. Experimental Results

This section evaluates the effectiveness of our developed method on eighteen benchmark datasets. In addition, we compare it to four other FS algorithms.

A. Dataset Description and Parameter Setting

To verify the performance of our proposed FS Wrapper model which is based on the AOA algorithm, we performed a comparison of our method with the following FS optimization algorithms: SCA, WOA, BAO and MFO. Each solution was evaluated using a fitness function given in equation (Eq. 15), which improves the accuracy of predictions and reduces the number of features. The effectiveness of the approaches proposed in the experiments is measured by several well-known and widely used evaluation metrics in this field [52], [53] which are cited in the next section. All algorithms are hybrid with the standard KNN machine learning classifier (with k=5). All experiments are conducted under the same conditions and the average of each evaluation metric is calculated from 20 independent runs. The proposed algorithms were tested on sixteen benchmark datasets (UCI repository) [54] to discuss the experimental results and comparisons. The details of the datasets are represented in Table I. This dataset is divided into 80% training and 20% testing. All parameters of the algorithms used in this paper are given in Table II with the following common parameters: N=10 (the number of search agents) and T=80 (the maximum number of iterations). In our study we used the MATLAB 2014b platform under the Microsoft Windows 10 professional 64bit operating system with the following hardware configuration: an Intel®Core TM i5 processor (3.20 GHz) with 4 GB RAM.

TABLE I. The Characteristics of the Dataset

No	Dataset	No of features	No of instances
DS1	AA	30	102
DS2	BreastEW	30	596
DS3	CongressEW	16	435
DS4	Exactly	13	1000
DS5	Exactly2	13	1000
DS6	HeartEW	13	270
DS7	IonosphereEW	34	351
DS8	KrvskpEW	36	3196
DS9	M-of-n	13	1000
DS10	penglungEw	325	569
DS11	sonarEW	60	208
DS12	SpecEw	22	267
DS13	Vote	16	300
DS14	WaveformEW	40	5000
DS15	WineEW	13	178
DS16	Zoo	16	101

TABLE II. Parameters of the Algorithms Used in the Experiments

Algorithm	Parameter	Value
SCA	a	2
WOA	$\vec{a}$	decreases linearly from 2 to 0
	$\vec{a}_2$	decreases linearly from −1 to −2
	b	1
BAO	a	0.8
MFO	b	0.75
AOA	c_1	2
	c_2	6
	α	0.9
	β	0.1

B. Performance Evaluation Measures

The effectiveness of the proposed method was evaluated by well-known and widely used evaluation metrics [49] which are:

- Accuracy is the proportion of the number of positive tuples and negative tuples obtained by the classification algorithms in the total number of hits, as shown in Equation 16.

$$Acc = \frac{(tP+tN)}{(tP+tN+fP+fN)} \tag{16}$$

tP, tN, fP and fN denote the numbers of true positives, true negatives, false positives and false negatives, respectively.

- F-score (FScore) - This is an evaluation of the precision of the classifier, statistically, it represents the harmonic mean between recall and accuracy. The formula for F-score is given by the following expression:

$$F_{score} = 2 \times \frac{precision \times recall}{precision + recall} = \frac{tP}{tP+\frac{1}{2}(fP+fN)} \tag{17}$$

- The average (MEAN)- The average of the fitness function values obtained by running an optimization algorithm M several times. Equation 18 displays the mathematical expression of the average of the fitness values:

$$MEAN = \frac{1}{M}\sum_{i=1}^{M} fit_{\vartheta} \tag{18}$$

- The best and worst fitness values are represented in equations 19 and 20 respectively:

$$BEST = Min_{i=1}^{M} fit_{\vartheta} \tag{19}$$

$$WORST = Max_{i=1}^{M} fit_{\vartheta} \tag{20}$$

- Standard Deviation (StD)

The standard deviation (StD) represents the variance of the best solutions found for each optimization algorithm run for M times. The StD is given in equation 21:

$$StD = \sqrt{\frac{1}{M-1}\sum_{i=1}^{M}(fit_{\vartheta} - MEAN)^2} \tag{21}$$

- Average selection size (AVRSS)

The average selection size is the average size of the selected features relative to the total number of features. This measure is as in equation (22).

$$AVRSS = \frac{1}{M}\sum_{i=1}^{M} \frac{size(fit_{\vartheta})}{D} \tag{22}$$

Where size(x) is the number of values of vector x for each run *i*, and *D* is the number of features in the original data set.

C. Results and Discussion

In all experiments, we use the cross-validation process to evaluate the performance. The k-fold cross validation algorithm consists in splitting the initial set of examples D into k blocks. We then repeat k evaluation learning phases, where a hypothesis h is obtained by learning on (k-1) blocks of data and tested on the remaining block. The error estimator is obtained as the average of the k empirical errors thus obtained.

In this section, as shown in Tables (III, IV, V, VI, VII, VIII) and Figures (3, 4), we evaluate the performance of the proposed method using various feature selection sets and compare it with other algorithms to prove its capability. The bold values in each row indicate the best result among the five algorithms.

Table III shows the best results of fitness values. In this table, we can notice that our proposed method presented competitive results with other comparison algorithms. AOA obtained the best results in nine datasets (i.e., DS2, DS3, DS8, DS10, DS11, DS12, DS13 and DS16), while in the other remaining datasets, the best results of the fitness values are not stable in one algorithm, but overall, there is not a big difference compared to the values obtained by our method in all test of the experiment.

TABLE III. The Best Fitness Values Obtained by the Different Optimizers

dataset	SCA	WOA	BAO	MFO	AOA
DS1	0,01293	0,09771	0,06521	0,08923	**0,06265**
DS2	0,06101	0,07923	0,08452	0,12002	**0,05567**
DS3	0,11321	0,10232	0,11082	0,07965	**0,03512**
DS4	0,05934	0,27624	**0,05385**	0,07212	0,05694
DS5	0,21361	**0,21331**	0,22971	0,24012	0,22968
DS6	0,43210	0,31812	0,28721	**0,24324**	0,26993
DS7	**0,15349**	0,27305	0,17434	0,21871	0,15422
DS8	0,21320	0,11325	0,08786	0,10239	**0,07923**
DS9	**0,06134**	0,24533	0,09742	0,11552	0,07038
DS10	0,06834	0,08972	0,08191	0,08321	**0,05396**
DS11	0,05940	0,09308	0,08229	0,23075	**0,00601**
DS12	0,07941	0,23931	0,20487	0,09501	**0,05729**
DS13	0,21431	0,32945	0,28214	0,20815	**0,20726**
DS14	0,31401	0,09329	0,10051	**0,24883**	0,29627
DS15	**0,02546**	0,08757	0,05384	0,03021	0,05388
DS16	0,02539	0,08611	0,05690	0,02612	**0,02034**

On the other hand, for the worst results of fitness values, as shown in Table IV, our proposed method is superior to other algorithms. AOA achieved the best results in 50% of all datasets used in the experiment (i.e., DS1, D23, DS3, DS4, DS9, DS10, DS13 and DS16). WOA achieved second place in 19% of all datasets (i.e., DS7, DS11, and DS12), followed by BAO and MFO with 13% each; while the SCA algorithm was ranked as the least successful.

The results of feature selection of all methods using the average of fitness values are recorded in Table V. We can observe from this table that our proposed technique got the best average in nine out of 16 datasets (i.e., 56%), so it was ranked first. AOA got the second place in four of the 16 datasets (i.e., DS1, DS8, DS10 and D14). BAO had the worst performance.

Table VI shows the StD values of the compared algorithms. It can be noticed that the proposed method based on the AOA algorithm has an excellent performance. The StD values prove that the proposed AOA is a powerful method for solving feature selection problems. AOA got more than 56% of the best cases according to the StD value, followed by SCA, BAO and MFO which got almost 13% for each and finally WOA algorithm is ranked as the bad method according to the StD values in the dataset compared to other methods. But, in general, the average StD values for all the datasets used in our experiment selected the performance of the algorithms even if they did not get the best StD value, for example in the case of WOA algorithm.

TABLE IV. Worst Fitness Values Obtained by the Different Optimizers

dataset	SCA	WOA	BAO	MFO	AOA
DS1	0,0452	0,0597	0,0464	0,0543	**0,0417**
DS2	0,0638	0,0492	0,0646	0,1107	**0,0562**
DS3	0,1774	0,0563	0,0762	0,0767	**0,0216**
DS4	0,0619	0,0463	0,0538	0,0617	**0,0321**
DS5	**0,2769**	0,2799	0,2778	0,2899	0,2957
DS6	0,4853	0,2101	0,2342	**0,2033**	0,2136
DS7	0,3038	**0,1231**	0,1193	0,1693	0,1345
DS8	0,2538	0,0743	**0,0435**	0,1048	0,0753
DS9	0,1613	0,0959	0,0859	0,1001	**0,0764**
DS10	0,0736	0,0616	0,0513	0,0615	**0,0462**
DS11	0,2007	**0,0030**	0,0525	0,1992	0,0041
DS12	0,1535	**0,0481**	0,1045	0,0750	0,0553
DS13	0,2703	0,2184	0,2182	0,1781	**0,1694**
DS14	0,3756	0,0637	**0,0529**	0,2210	0,2213
DS15	0,1433	0,0692	0,0385	**0,0275**	0,0538
DS16	0,1453	0,0437	0,0319	0,0153	**0,0100**

TABLE V. The Average Fitness Values Obtained by the Different Optimizers

dataset	SCA	WOA	BAO	MFO	AOA
DS1	**0,0291**	0,0787	0,0558	0,0718	0,0522
DS2	0,0624	0,0642	0,0746	0,1154	**0,0559**
DS3	0,1453	0,0793	0,0935	0,0782	**0,0284**
DS4	0,0606	0,1613	0,0538	0,0669	**0,0445**
DS5	**0,2452**	0,2466	0,2537	0,2649	0,2626
DS6	0,4587	**0,2241**	0,2607	0,2233	0,2418
DS7	0,2286	0,2231	0,1468	0,1940	**0,1444**
DS8	0,2335	0,0938	**0,0457**	0,1036	0,0773
DS9	0,1113	0,1556	0,0917	0,1078	**0,0734**
DS10	0,0710	**0,0457**	0,0666	0,0724	0,0501
DS11	0,1300	0,0481	0,0674	0,2150	**0,0051**
DS12	0,1164	0,1437	0,1547	0,0850	**0,0563**
DS13	0,2423	0,2739	0,2502	0,1931	**0,1883**
DS14	0,3448	0,0785	0,0767	**0,2349**	0,2588
DS15	0,0844	0,0784	0,0462	**0,0289**	0,0539
DS16	0,0854	0,0649	0,0444	0,0207	**0,0152**

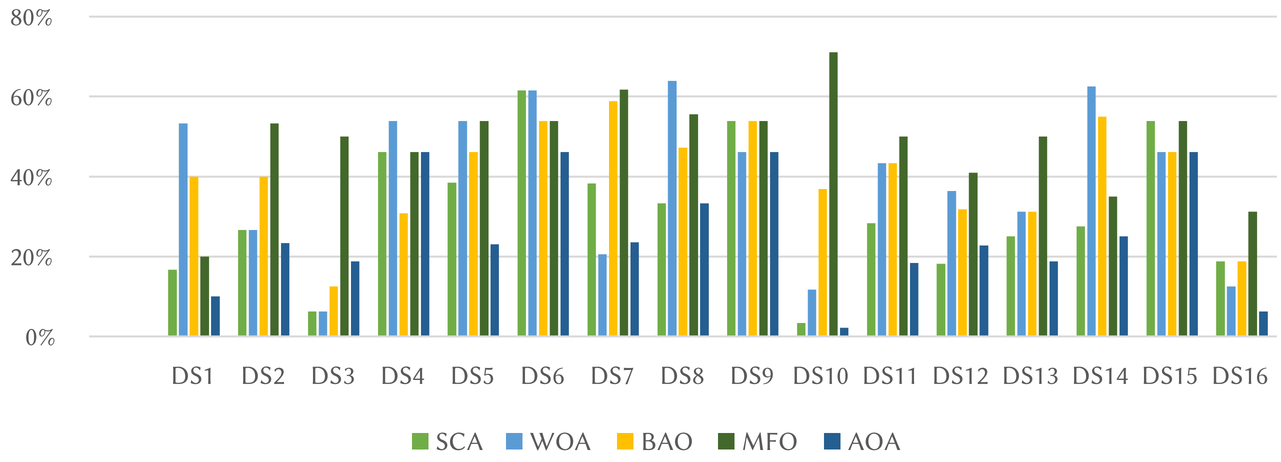


Fig 3. The percentage of the selected features for the comparative methods.

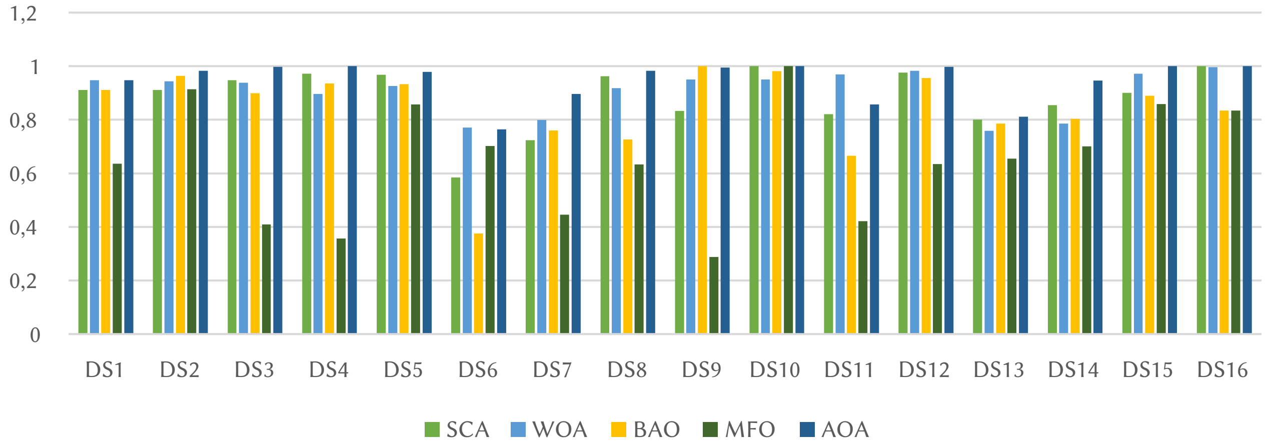


Fig 4. Comparison between the different proposed methods in terms of classification accuracy on all datasets.

TABLE VI. The Average Fitness Values Obtained by the Different Optimizers

dataset	SCA	WOA	BAO	MFO	AOA
DS1	0,0064	0,0038	0,0065	0,0097	**0,0013**
DS2	0,0071	0,0039	0,0037	0,0032	**0,0113**
DS3	**0,0021**	0,0041	0,0078	0,0078	0,0029
DS4	0,0021	0,0063	0,0052	0,0077	**0,0000**
DS5	0,0357	0,0057	0,0042	0,0072	**0,0012**
DS6	0,0056	0,0083	0,0276	**0,0102**	0,0155
DS7	0,0131	0,0074	0,0081	0,0331	**0,0126**
DS8	0,0036	**0,0031**	0,0098	0,0393	0,0085
DS9	0,0106	0,0031	**0,0000**	0,0141	0,0074
DS10	**0,0000**	0,0106	0,0022	0,0133	0,0034
DS11	0,0058	0,0064	0,0134	**0,0000**	0,0054
DS12	0,0027	0,0081	0,0052	0,0134	**0,0000**
DS13	0,0098	0,0092	0,0093	0,0112	**0,0087**
DS14	0,0089	0,0087	**0,0035**	0,0052	0,0142
DS15	0,0036	0,0042	0,0034	0,0091	**0,0000**
DS16	0,0000	0,0002	0,0031	0,0063	**0,0000**

Moreover, the percentage of the best number of selected features is presented in Fig. 3. The performance of our proposed method is clearly visible in this figure. To obtain higher accuracy values, it is necessary to obtain a small number of selected features. It can be seen that WOA obtained the smallest number of features in 68% of the data sets. WOA obtained the second largest and smallest number of features in 22% of the datasets, followed by SCA; while MFO recorded the largest number of features among all algorithms.

From Fig. 4, we notice that the results of the envelope attribute selection method based on the AOA algorithm combined with the KNN classifier are better than the results obtained by the other methods with the same classifier in terms of accuracy. The results reach a classification accuracy of 97% for the average of all datasets by our approach, followed by the optimization by WOA which has a classification accuracy of 91%, while MFO showed the lowest accuracy in the average of all datasets. We also note that the results reach 100% classification rate for four datasets used in the experiments, namely DS4, DS10, DS15 and DS16.

To statistically validate our study and to give it more value, we applied the Wilcoxon statistical test [55]. This test (Wilcoxon Rank-sum) is a non-parametric statistical test that tests the hypothesis that the medians of each of two groups of data are close. As in all statistical tests, it allows to accept and also to reject the NULL hypothesis. The latter considers that the median of two real data vectors X and Y is fair. The p-value was compared at a significance level of 0.05. For ease of understanding, the symbols "w/t/l" indicate that the AOA is superior (win), equal (tie) and inferior (lose) to the other algorithms. So as shown by the p-values presented in Table VII, we can see that the method proposed by AOA brings a significant improvement over the algorithms: SCA and WOA in most of the collections used. However, this superiority is statistically weak for the other algorithms such as BAO and MFO. Therefore, AOA shows a good performance in terms of Wilcoxon test and it can be chosen as a reference algorithm as it offers significant classification results in this work. Also, in terms of F-score values obtained, the proposed method outperforms the other algorithms on thirteen datasets as shown in Table VIII, on the other hand the SCA and WOA techniques are only effective on two datasets.

TABLE VII. P-Values of the Wilcoxon Rank Sum Test of the AOA Accuracy Results Versus Other Algorithms

dataset	p-values			
	SCA	WOA	BAO	MFO
DS1	0,0064304	0,003800	0,066050	0,078097
DS2	0,0072970	0,003399	0,003027	0,003002
DS3	0,0021002	0,104111	0,200778	0,120078
DS4	0,0036721	0,006333	0,000102	0,008077
DS5	0,0354567	0,005713	0,010142	0,080172
DS6	0,0056000	0,008314	0,027036	0,109102
DS7	0,0105631	0,067401	0,080081	0,321031
DS8	0,0030116	0,003120	0,002098	0,030093
DS9	0,0141206	0,003100	0,000000	0,000000
DS10	0,0000000	0,010612	0,070022	0,013073
DS11	0,0059018	0,006224	0,007134	0,000056
DS12	0,0028900	0,208100	0,005332	0,098034
DS13	0,0049098	0,002340	0,070293	0,011002
DS14	0,0030989	0,078712	0,003505	0,108052
DS15	0,0036003	0,064233	0,201034	0,300091
DS16	0,0000000	0,000400	0,000305	0,003005
w|t|l	16|0|0	11|5|0	9|7|0	8|8|0

TABLE VIII. The F-Score Values Obtained by the Different Optimizers

dataset	SCA	WOA	BAO	MFO	AOA
DS1	0,9201	0,9555	0,9202	0,7365	**0,9578**
DS2	0,9145	0,9541	0,9661	0,9221	**0,9896**
DS3	0,9565	0,9478	0,9093	0,5393	**0,9979**
DS4	0,9773	0,9160	0,9398	0,4571	**1,0000**
DS5	0,9731	0,9321	0,9422	0,8987	**0,9833**
DS6	0,6653	**0,7921**	0,5342	0,7201	0,7788
DS7	0,7531	0,8021	0,7696	0,5261	**0,9038**
DS8	0,9679	0,9285	0,7337	0,7228	**0,9902**
DS9	0,8435	0,9555	**1,0000**	0,4101	0,9995
DS10	**1,0000**	0,9517	0,9899	**1,0000**	**1,0000**
DS11	0,8621	**0,9722**	0,7156	0,5210	0,8621
DS12	0,9775	0,9866	0,9663	0,6896	**0,9989**
DS13	0,8101	0,7623	0,7899	0,7034	**0,8811**
DS14	0,8651	0,7933	0,8132	0,7901	**0,9543**
DS15	0,9102	0,9798	0,9067	0,8991	**1,0000**
DS16	**1,0000**	0,9998	0,8446	0,8436	**1,0000**

Looking at the tables from III to VIII, figures from 3 to 4 and the in-depth analysis of the results of the developed method against the other comparison methods, we can see that the AOA algorithm achieved the best results in terms of accuracy, number of selected features and fitness value on the majority of the datasets. The results that are displayed in the previous tables show that the effectiveness of each algorithm depends on the dataset used, but overall and according to the results of the average values of the evaluation criteria used in this paper on the dataset, we can therefore conclude that our approach often has superiority over the other methods based on WOA, SCA, MFO and BAO algorithms that are performed under the same conditions and on the same datasets that we have mentioned previously.

In general, the previous results show that there is a significant improvement in solving feature selection problems using the operators of the AOA algorithm. Therefore, it can be said that AOA can be considered as an effective optimization algorithm, especially for solving feature selection problems.

V. Conclusion

In this work, we have addressed the feature selection problem by using a recent optimizer called the Archimedes' optimization algorithm (AOA). Experiments are applied on sixteen different datasets (UCI) to handle the FS optimization task and to study the effectiveness of the proposed method. The latter has been compared by four FS methods based on WOA, SCA, FMO and BAO algorithms with the same classifier which is KNN. We also used several evaluation criteria to properly assess different aspects of the performance of the compared algorithms. The comparisons revealed that our technique achieved the best average fitness in nine of the 16 datasets, the lowest StD value in 56% of the datasets, and the lowest feature count in 68% of the datasets. These results indicate that AOA is able to select a small number of features and achieve very high classification accuracy. Therefore, we conclude that the operators of the AOA algorithm improve the mining and exploration phases well which increases the efficiency of this algorithm to solve classification problems.

References

[1] M. A. Khan et al., "Cucumber leaf diseases recognition using multi level deep entropy-ELM feature selection," Applied Sciences, vol. 12, no. 2, p. 593, 2022.

[2] M. A. Khan et al., "A fused heterogeneous deep neural network and robust feature selection framework for human actions recognition," Arabian Journal for Science and Engineering, pp. 1–16, 2021.

[3] A. Mehmood, U. Tariq, C. W. Jeong, Y. Nam, R. R. Mostafa, and A. Elaeiny, "Human Gait Recognition: A Deep Learning and Best Feature Selection Framework," Computers, Materials & Continua, vol. 70, pp. 343–360, 2022.

[4] N. Hussain et al., "Multiclass cucumber leaf diseases recognition using best feature selection," Computers, Materials & Continua, vol. 70, pp. 3281–3294, 2022.

[5] F. Zia et al., "A multilevel deep feature selection framework for diabetic retinopathy image classification," 2022.

[6] A. Rehman, M. A. Khan, T. Saba, Z. Mehmood, U. Tariq, and N. Ayesha, "Microscopic brain tumor detection and classification using 3D CNN and feature selection architecture," Microscopy Research and Technique, vol. 84, no. 1, pp. 133–149, 2021.

[7] M. U. Khan et al., "Expert hypertension detection system featuring pulse plethysmograph signals and hybrid feature selection and reduction scheme," Sensors, vol. 21, no. 1, p. 247, 2021.

[8] M. A. Khan et al., "Multimodal brain tumor classification using deep learning and robust feature selection: A machine learning application for radiologists," Diagnostics, vol. 10, no. 8, p. 565, 2020.

[9] L. Khrissi, H. Satori, K. Satori, and N. el Akkad, "An Efficient Image Clustering Technique based on Fuzzy C-means and Cuckoo Search Algorithm," International Journal of Advanced Computer Science and Applications, vol. 12, no. 6, pp. 423–432, 2021, doi: 10.14569/IJACSA.2021.0120647.

[10] L. Khrissi, N. E. Akkad, H. Satori, and K. Satori, "Color image segmentation based on hybridization between Canny and k-means," in 7th Mediterranean Congress of Telecommunications 2019, CMT 2019, 2019. doi: 10.1109/CMT.2019.8931358.

[11] D. Yousri, M. Abd Elaziz, L. Abualigah, D. Oliva, M. A. A. Al-Qaness, and A. A. Ewees, "COVID-19 X-ray images classification based on enhanced fractional-order cuckoo search optimizer using heavy-tailed distributions," Applied Software Computing, vol. 101, p. 107052, 2021.

[12] Z. Faska, L. Khrissi, K. Haddouch, and N. el Akkad, "A Powerful and Efficient Method of Image Segmentation Based on Random Forest Algorithm," in Digital Technologies and Applications, 2021, pp. 893–903.

[13] L. Khrissi, N. El Akkad, H. Satori, and K. Satori, "Simple and Efficient Clustering Approach Based on Cuckoo Search Algorithm," 2020 Fourth International Conference On Intelligent Computing in Data Sciences (ICDS), pp. 1–6, Oct. 2020, doi: 10.1109/ICDS50568.2020.9268754.

[14] S. Cheng, L. Ma, H. Lu, X. Lei, and Y. Shi, "Evolutionary computation for solving search-based data analytics problems," Artificial Intelligence Review, vol. 54, no. 2, pp. 1321–1348, 2021, doi: 10.1007/s10462-020-09882-x.

[15] D. S. A. Elminaam, S. A. Ibrahim, E. H. Houssein, and S. M. Elsayed, "An Efficient Chaotic Gradient-Based Optimizer for Feature Selection," IEEE Access, vol. 10, pp. 9271–9286, 2022, doi: 10.1109/ACCESS.2022.3143802.

[16] H. Moussaoui, N. el Akkad, and M. Benslimane, "Moroccan Carpets Classification Based on SVM Classifier and ORB Features," in International Conference on Digital Technologies and Applications, 2022, pp. 446–455.

[17] C. C. Aggarwal, X. Kong, Q. Gu, J. Han, and P. S. Yu, "Active learning: A survey," Data Classification: Algorithms and Applications, pp. 571–605, 2014, doi: 10.1201/b17320.

[18] E. H. Houssein, A. G. Gad, Y. M. Wazery, and P. N. Suganthan, "Task scheduling in cloud computing based on meta-heuristics: Review, taxonomy, open challenges, and future trends," Swarm and Evolutionary Computation, vol. 62, p. 100841, 2021.

[19] F. A. Hashim, E. H. Houssein, K. Hussain, M. S. Mabrouk, and W. Al-Atabany, "A modified Henry gas solubility optimization for solving motif discovery problem," Neural Computing & Applicationsvol. 32, no. 14, pp. 10759–10771, 2020.

[20] N. Neggaz, E. H. Houssein, and K. Hussain, "An efficient henry gas solubility optimization for feature selection," Expert Systems With Applications, vol. 152, p. 113364, 2020.

[21] L. Khrissi, N. el Akkad, H. Satori, and K. Satori, "Image Segmentation Based on K-means and Genetic Algorithms," in Advances in Intelligent Systems and Computing, 2020, vol. 1076, pp. 489–497. doi: 10.1007/978-981-15-0947-6_46.

[22] L. Khrissi, N. el Akkad, H. Satori, and K. Satori, “Clustering method and sine cosine algorithm for image segmentation,” Evolutionary Intelligence, Jan. 2021, doi: 10.1007/s12065-020-00544-z.

[23] L. Khrissi, N. el Akkad, H. Satori, and K. Satori, “A Performant Clustering Approach Based on An Improved Sine Cosine Algorithm,” International Journal of Computing, pp. 159–168, Jun. 2022, doi: 10.47839/ijc.21.2.2584.

[24] M. Merras, N. el Akkad, A. Saaidi, A. G. Nazih, and K. Satori, “Camera calibration with varying parameters based on improved genetic algorithm,” WSEAS Transactions on Computersvol. 13, pp. 129–137, 2014.

[25] N. el Akkad, M. Merras, A. Saaidi, and K. Satori, “Robust method for self-calibration of cameras having the varying intrinsic parameters,” Journal of Theoretical and Applied Information Technology, vol. 50, no. 1, pp. 57 – 67. 2013.

[26] H. M. Zawbaa, E. Emary, and B. Parv, “Feature selection based on antlion optimization algorithm,” Proceedings of 2015 IEEE World Conference on Complex Systems, WCCS 2015, 2016, doi: 10.1109/ICoCS.2015.7483317.

[27] M. Mafarja et al., “Binary dragonfly optimization for feature selection using time-varying transfer functions,” 2018, doi: 10.1016/j.knosys.2018.08.003.

[28] A. G. Hussien, A. E. Hassanien, E. H. Houssein, S. Bhattacharyya, and M. Amin, “S-shaped binary whale optimization algorithm for feature selection,” vol. 727. Springer Singapore, 2019. doi: 10.1007/978-981-10-8863-6_9.

[29] A. T. Sahlol, D. Yousri, A. A. Ewees, M. A. A. Al-Qaness, R. Damasevicius, and M. A. Elaziz, “COVID-19 image classification using deep features and fractional-order marine predators algorithm,” Scientific Reports , vol. 10, no. 1, pp. 1–15, 2020.

[30] A. I. Hafez, H. M. Zawbaa, E. Emary, and A. E. Hassanien, “Sine cosine optimization algorithm for feature selection,” in 2016 international symposium on innovations in intelligent systems and applications (INISTA), 2016, pp. 1–5.

[31] P. C. Chiu, A. Selamat, O. Krejcar, K. K. Kuok, E. Herrera-Viedma, and G. Fenza, “Imputation of Rainfall Data Using the Sine Cosine Function Fitting Neural Network.,” International Journal of Interactive Multimedia & Artificial Intelligence, vol. 6, no. 7, 2021.

[32] Y. Zhang, R. Liu, X. Wang, H. Chen, and C. Li, “Boosted binary Harris hawks optimizer and feature selection,” Engineering with Computers, vol. 37, no. 4, pp. 3741–3770, 2021.

[33] N. Neggaz, A. A. Ewees, M. Abd Elaziz, and M. Mafarja, “Boosting salp swarm algorithm by sine cosine algorithm and disrupt operator for feature selection,” Expert Systems With Applications, vol. 145, p. 113103, 2020.

[34] M. M. Mafarja and S. Mirjalili, “Hybrid Whale Optimization Algorithm with simulated annealing for feature selection,” Neurocomputing, vol. 260, pp. 302–312, Oct. 2017, doi: 10.1016/j.neucom.2017.04.053.

[35] M. Abdel-Basset, W. Ding, and D. El-Shahat, “A hybrid Harris Hawks optimization algorithm with simulated annealing for feature selection,” Artificial Intelligence Review, vol. 54, no. 1, pp. 593–637, 2021.

[36] F. A. Hashim, K. Hussain, E. H. Houssein, M. S. Mabrouk, and W. Al-Atabany, “Archimedes optimization algorithm: a new metaheuristic algorithm for solving optimization problems,” Applied Intelligence, vol. 51, no. 3, pp. 1531–1551, 2021, doi: 10.1007/s10489-020-01893-z.

[37] I. Neggaz and H. Fizazi, “An Intelligent handcrafted feature selection using Archimedes optimization algorithm for facial analysis,” Soft Computing, vol. 26, pp. 10435–10464, 2022.

[38] V. Janamala and K. Radha Rani, “Optimal allocation of solar photovoltaic distributed generation in electrical distribution networks using Archimedes optimization algorithm,” Clean Energy, vol. 6, no. 2, pp. 271–287, 2022.

[39] L. Zhang, J. Wang, X. Niu, and Z. Liu, “Ensemble wind speed forecasting with multi-objective Archimedes optimization algorithm and sub-model selection,” Applied Energy, vol. 301, p. 117449, 2021.

[40] R. A. Khan et al., “Archimedes Optimization Algorithm Based Selective Harmonic Elimination in a Cascaded H-Bridge Multilevel Inverter,” Sustainability, vol. 14, no. 1, p. 310, 2021.

[41] A. S. Desuky, S. Hussain, S. Kausar, M. A. Islam, and L. M. el Bakrawy, “EAOA: An Enhanced Archimedes Optimization Algorithm for Feature Selection in Classification,” IEEE Access, vol. 9, pp. 120795–120814, 2021.

[42] E. H. Houssein, B. E. Helmy, H. Rezk, and A. M. Nassef, “An enhanced Archimedes optimization algorithm based on Local escaping operator and Orthogonal learning for PEM fuel cell parameter identification,” Engineering Applications of Artificial Intelligence, vol. 103, p. 104309, 2021, doi: https://doi.org/10.1016/j.engappai.2021.104309.

[43] O. Akdag, “A Improved Archimedes Optimization Algorithm for multi/single-objective Optimal Power Flow,” Electric Power Systems Research, vol. 206, p. 107796, 2022.

[44] D. Izzo, M. Märtens, and B. Pan, “A survey on artificial intelligence trends in spacecraft guidance dynamics and control,” Astrodynamics, vol. 3, no. 4, pp. 287–299, 2019, doi: 10.1007/s42064-018-0053-6.

[45] A. A. Ewees, M. A. el Aziz, and A. E. Hassanien, “Chaotic multi-verse optimizer-based feature selection,” Neural Computing and Applications, vol. 31, no. 4, pp. 991–1006, 2019, doi: 10.1007/s00521-017-3131-4.

[46] Y. Zhang, R. Liu, X. Wang, H. Chen, and C. Li, “Boosted binary Harris hawks optimizer and feature selection,” Engineering with Computers, vol. 37, no. 4, pp. 3741–3770, 2021.

[47] G. Tikhe, T. Joshi, A. Lahorkar, A. Sane, and J. Valadi, “Feature selection using equilibrium optimizer,” in Data Engineering and Intelligent Computing, Springer, 2021, pp. 307–315.

[48] C. Rorres, “Completing book II of Archimedes’s on floating bodies,” The mathematical intelligencer, vol. 26, no. 3, pp. 32–42, 2004.

[49] Y. Y. Yiming, “An Evaluation of Statistical Approaches to Text Categorization,” Journal of Information Retrieval, vol. 1, pp. 67–88, 1999.

[50] M. Mafarja, I. Aljarah, H. Faris, A. I. Hammouri, A. M. Al-Zoubi, and S. Mirjalili, “Binary grasshopper optimisation algorithm approaches for feature selection problems,” Expert Systems with Applications, vol. 117, pp. 267–286, 2019, doi: 10.1016/j.eswa.2018.09.015.

[51] T. Thaher, A. A. Heidari, M. Mafarja, J. S. Dong, and S. Mirjalili, “Binary Harris Hawks Optimizer for High-Dimensional, Low Sample Size Feature Selection,” in Evolutionary Machine Learning Techniques: Algorithms and Applications, S. Mirjalili, H. Faris, and I. Aljarah, Eds. Singapore: Springer Singapore, 2020, pp. 251–272. doi: 10.1007/978-981-32-9990-0_12.

[52] A. S. Desuky and L. M. el Bakrawy, “Improved prediction of post-operative life expectancy after thoracic surgery,” Advances in Systems Science and Application, vol. 16, no. 2, pp. 70–80, 2016.

[53] M. Gong, “A Novel Performance Measure for Machine Learning Classification,” International Journal of Managing Information Technology, vol. 13, no. 1, pp. 11–19, 2021, doi: 10.5121/ijmit.2021.13101.

[54] D. Dua and C. Graff, “UCI machine learning repository,” 2017.

[55] D. Rey and M. Neuhäuser, “Wilcoxon-signed-rank test,” in International encyclopedia of statistical science, Springer, 2011, pp. 1658–1659.

Lahbib Khrissi

Lahbib Khrissi received the bachelor’s and master’s degrees from SMBA-Fez University in 2000 and 2007 respectively. He is currently working toward the PhD degree in the LISAC Laboratory at faculty of Sciences, Sidi Mohammed Ben Abdellah University, Fez, Morocco. His current research interests include Data mining, Image processing and Artificial intelligence.

Nabil El Akkad

Nabil El Akkad received the PhD degree from SMBA-Fez University in 2014. He is currently a professor of computer science at National School of Applied Sciences (ENSA) of Fez, Sidi Mohammed Ben Abdellah University, Fez, Morocco. He is a member of the LISA and LIIAN Laboratories. His research interests include Artificial intelligence, Image processing, Data mining, Camera self calibration, 3D reconstruction, Machine learning, Real-time rendering and Cryptography.

Hassan Satori

Hassan Satori is Associate Professor, Department of Computer Science, Faculty of Sciences, Dhar El Mahraz, Sidi Mohammed Ben Abdellah University -Fez- Morocco. He is Ph.D in Speech recognition and Signal Processing in 2009. He received a MSc in Physics from the Mohamed Premier University, and a Ph.D in Nanotechnology and computer simulation, also from the same University, in 2001. From 2001-2002 he was a postdoctoral fellow in Physics, mathematical modeling and computer simulation at the Chemnitz University of Technology, Germany, and was then a postdoctoral fellow in computer simulation at Department of Interface Chemistry and Surface Engineering, Max-Planck-Institut, Düsseldorf, Germany, (2003-2005). Dr. Hassan SATORI has large academic teaching and research experience in the fields.

Khalid Satori

Khalid Satori received the PhD degree from the National Institute for the Applied Sciences INSA at Lyon in 1993. He is currently a professor of computer science at SMBA-Fez University. He is the director of the LISAC Laboratory. His research interests include real-time rendering, Image-based rendering, virtual reality, biomedical signal, camera self calibration, genetic algorithms and 3D reconstruction.

Regular Issue

Machine Learning Based Agricultural Profitability Recommendation Systems: A Paradigm Shift in Crop Cultivation

Nilesh P. Sable[1], Rajkumar V. Patil[2], Mahendra Deore[3], Ratnmala Bhimanpallewar[4], Parikshit N. Mahalle[5] *

[1] Department of Computer Science & Engineering (Artificial Intelligence), Bansilal Ramnath Agarwal Charitable Trust's Vishwakarma Institute of Information Technology, Pune (India)
[2] MIT Art, Design & Technology University, Pune (India)
[3] Department of Computer Engineering, MKSSS's Cummins College of Engineering for Women, Pune-411052, Maharashtra, (India)
[4] Department of Information Technology, Bansilal Ramnath Agarwal Charitable Trust's Vishwakarma Institute of Information Technology, Pune (India)
[5] Professor and Dean R&D, Bansilal Ramnath Agarwal Charitable Trust's, Vishwakarma Institute of Technology, Pune (India)

* Corresponding author: drsablenilesh@gmail.com (N. P. Sable), rajkumar.v.patil30@gmail.com (R. V. Patil), mdeore83@gmail.com (M. Deore), ratnmalab@gmail.com (R. Bhimanpallewar), aalborg.pnm@gmail.com (P. N. Mahalle)

Received 24 August 2023 | Accepted 16 July 2024 | Published 30 October 2024

unir
LA UNIVERSIDAD
EN INTERNET

Abstract

In India, the demand for fruits and vegetables has been consistently increasing alongside the rising population, making crop production a crucial aspect of agriculture. However, despite the growing demand and potential profitability, farmers have been slow to transition from traditional food grain crops to fruits and vegetables. In this paper, we explore the changing demands of food categories in India, highlighting the shift towards increased consumption of fruits and vegetables. Despite the potential benefits, farmers face various challenges and uncertainties associated with cultivating these crops. To address this, we propose the use of Machine Learning (ML) and Deep Learning (DL) techniques to analyze historical market price data for fruits and vegetables from 2016 to 2021 and predict future prices. This accurate prediction system will aid farmers in deciding which crops to grow and when to harvest, ultimately maximizing profits.

Keywords

Agriculture, Cultivation, Data Analysis, Machine Learning, Regression.

DOI: 10.9781/ijimai.2024.10.005

I. Introduction

AGRICULTURE, the world's oldest and most important sector, has always been essential for supplying food, fibers, and fuel to humanity. Archaeological evidence places the origins of farming at about 10,000 years ago, when people began to depend on it for their food [1]. Agriculture plays a crucial role in the development of civilizations by cultivating the soil and raising livestock. Nevertheless, throughout millennia, agricultural growth progressed gradually. Using fire to regulate plant development was a common practice in early agricultural techniques since people had seen how well-established vegetation was following wildfires. Farmers gradually started tilling the soil and producing crops on tiny pieces of land by hand using simple equipment. As time went on, productive farming implements were developed, and yield-boosting irrigation methods were mastered [2].

As per the statistics of Annual crop production in India since 2003-04 fruit and vegetable production keeps on increasing [3]. With the continuous growth in population and changing food consumption patterns in India, there is an increasing demand for fruits and vegetables. However, farmers have been hesitant to shift from traditional food grain crops to fruits and vegetables due to various reasons. This paper aims to provide a solution by using ML and DL algorithms to analyze historical market price data of Mumbai Agricultural Produce Market Committee (Mumbai APMC) and predict fruit and vegetable prices, assisting farmers in making informed decisions about crop selection and harvesting.

Changes in Food Consumption Patterns: According to Dr. Richa Govil of Ashoka India, there has been a noteworthy rise in the consumption of fruits and vegetables despite a minor decline in the consumption of cereals like wheat and rice. Farmers have been sluggish to adopt fruit and vegetable crops despite this change [4]. Farmers face a number of difficulties, some of which have been highlighted

Please cite this article as: N. P. Sable, R. V. Patil, M. Deore, R. Bhimanpallewar, P. N. Mahalle. Machine Learning Based Agricultural Profitability Recommendation Systems: A Paradigm Shift in Crop Cultivation, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 1, pp. 39-54, 2024, http://dx.doi.org/10.9781/ijimai.2024.10.005

by a survey of those in the agricultural industry [4]. Farmers are reluctant to grow fruits and vegetables for a variety of reasons. Firstly, compared to food grains, these crops are riskier to grow since they are more reliant on environmental factors. The cultivation of fruits and vegetables also takes more labour, and developing nations like India have difficulties due to the lack of automation. There is necessity of a reliable prediction system. We suggest using Machine Learning (ML) and Deep Learning (DL) approaches to examine historical market price data for fruits and vegetables from 2016 to 2021 in order to address these issues. Farmers may decide which crops to cultivate and when to harvest them profitably by using prediction of future pricing for each month and year.

A. Key Highlight of Research

- Price Prediction for Fruits and Vegetables: The paper focuses on predicting market prices for fruits and vegetables to aid farmers in making informed selling decisions.
- A literature survey was done to analyze a wide range of methodologies for analyzing time series data, with a special focus on predicting fruit market prices.
- Analysis of ML and DL Algorithms: Conduct a comprehensive examination of ML and DL techniques that are applicable to predicting fruit market prices.
- Analysis of Regression Algorithms: An analysis of several regression methods and their mathematical models to determine their suitability for predicting prices.
- Evaluation Metrics: Discussion on diverse evaluation metrics applicable to regression algorithms, accompanied by their respective mathematical formulations.
- Determining the Optimal Regression Model: Applying evaluation metrics to identify the optimal regression model for precise price prediction.
- Predicting the optimal month for farmers to harvest their crops, maximizing the possible selling price in the market.

The paper is structured as follows: Section II provides a comprehensive literature survey on Time Series Data Analysis, as well as on ML and DL based approaches in regression algorithms. Section III outlines the Experimental Methodology employed in this study. In Section IV, we discuss the results obtained from our experiments Also we discuss future work and limitations of research. Finally, in Section V, we conclude the study and present future scope for further research.

II. Literature Survey

The agriculture industry as well as the economy at large heavily rely on predictions of fruit and vegetable prices in market yards. Farmers, sellers, and policymakers may make informed decisions about production, distribution, and marketing strategies due to accurate price projections. To predict the pricing of fruits and vegetables in market yards, researchers have used a variety of strategies and procedures throughout the years. This review of the literature seeks to provide readers an overview of the current research and methodology used to predict the prices of fruits and vegetables. Researcher used Time Series Analysis, ML and Artificial Intelligence (AI), Data Mining and Big Data Analytics. For accurate prediction, one must carefully examine historical data and take into account a number of variables that have an impact on market dynamics. To tackle this problem, researchers have used a variety of methods, such as time series analysis, ML, and data mining. Each methodology has its strengths and limitations, and the choice of a technique depends on the available data, research objectives, and computational resources.

A. Literature Survey on Time Series Analysis

Villaren M. Vibas et al. concentrated on the development of a mathematical model to analyze the retail price changes of essential agricultural commodities, particularly fruits (mango and banana) and vegetables (cabbage, pechay and tomato) in the Philippines' National Capital Region (NCR) [5]. The Philippine Statistics Authority (PSA) provided the study with data that covered the ten-year period from 2009 to 2018. The data was analyzed and predictive models were created using time series modelling approaches as ARIMA, SARIMA, and ARIMAx. The study's conclusions showed that during the span of 10 years, the monthly prices of all the items under investigation had increased. When projecting monthly retail prices of fruit commodities, the ARIMAX (5, 2, 2, x=mango) model was shown to be the most accurate, whereas the ARIMAX (2, 2, 1, x=banana) model excelled for bananas. The study suggested utilizing the ARIMAX (3, 2, 1, x=pechay) model for cabbage, the SARIMA (1, 1, 1)(1, 1, 1)12 model for pechay, and the SARIMA (2, 1, 1)(2, 1, 1)12 model for tomatoes for calculating monthly prices for vegetable commodities. The aim is to help consumers, farmers, traders, business owners, and policymakers make wise decisions about economic issues and long-term planning involving basic agricultural commodities in the NCR region.

Sarker Rakhal, et al. examines the dynamics of price transmission in the Canadian orange and apple markets. The analysis makes use of orange and apple import and retail prices of each month from 1996 to 2017 [6]. The author examines the amount, direction, and speed of price transmission between the upstream (import) and downstream (retail) levels using co-integration and error correction modelling techniques. According to the results, both commodities' import and retail prices have a single, long-term relationship, with the import price having an impact on the retail price. Additionally, the findings show that apples have asymmetric price transmission, whereby the margin corrects more quickly when it is constricted than when it is expanded.

Ali Jahangir et al. carried out research to examine the pricing and arrival patterns of apple produced at Jammu's Narwal market [7]. The Directorate of Horticulture, Planning and Marketing in Narwal provided ten years' worth of monthly secondary data on apple pricing and arrivals (from 2007–08 to 2016–17). Linear regression was used by author to find insight and pattern for apple. Moreover, seasonal indices were computed to investigate the cyclical changes in company activity linked to the year cycle. The results showed a favorable trend in both apple pricing and arrivals, with an anticipated yearly rise of Rs. 220.06 per quintal and arrivals of 15,969.42 quintals of apples. The primary period for apple arrivals in Narwal market was from August through January. Prices for apples ranged from the lowest in April to the highest in August. The seasonal indices showed that apple arrivals peaked in October and peaked at their lowest in April, whereas the seasonal indicator for pricing peaked in August and was lowest in April. Although all above time series analysis studies offer useful information for agricultural market decision-making, none of them particularly address price prediction or recommend the ideal month for farmers to harvest their crops for the highest potential price.

B. Literature Survey on Machine Learning (ML) and Deep Learning (DL)

L. Nassar et al. compare the effectiveness of deep learning (DL) models for predicting the prices of fresh produce (FP) markets with statistical and conventional ML models [8]. Two datasets are used: one from a website that lists daily crop prices in Taiwanese marketplaces, and the other, for daily strawberry transactions over a seven-year period, comes from a confidential source in Canada. The findings demonstrate that traditional ML models perform better than statistical models like ARIMA. Gradient Boosting performs well among the ML models, although simple and compound DL models meet it.

Convolutional Long Short-Term Memory Recurrent Neural Network (CNN-LSTM) with attention, a compound DL model, performs the best and can predict FP prices up to three weeks in advance.

Ifeanyi Okwuchi discusses the difficult challenge of predicting Fresh Produce (FP) pricing, taking into account elements like the produce's limited shelf life, inability to be stored for extended periods of time, and outside impacts like weather and climate change [9]. The goal of the project is to build machine learning-based models for FP yield and price prediction, including both traditional and deep learning models. A variety of Californian data are used, including weather, strawberry output, farm-gate pricing, and store purchase prices. To evaluate the different prediction models, the author suggests a brand-new aggregated error metric (AGM) that incorporates mean absolute error, mean squared error, and R2 coefficient of determination. In order to further enhance the predictions, stacking ensemble approaches are used, such as voting regressor and stacking using Support Vector Regression (SVR).

Razat Agarwal used machine learning techniques to classify and predict fruit images using a large dataset. Five supervised learning models were created and evaluated for their effectiveness in identifying fruit, including Random Forest (RF), Naive Bayes, K-Nearest Neighbour (KNN), and Support Vector Machine (SVM). According to the experimental results, SVM performed better than the other methods for both large (95 fruits, 48905 images) and small (18 fruits, 8846 images) datasets. Additionally, it was observed that reducing the number of fruits i.e. small dataset negatively impacted the accuracy of each algorithm [10].

R. Dharavath and E. Khosla address the growing concern of inflation in India, particularly in the region of Bengaluru, Karnataka. The study's objective is to evaluate local fruit and vegetable prices and use seasonal ARIMA to predict future prices. Policymakers and people may prevent price rises by acting proactively by anticipating prices. Strategies might be developed to cut the price of fruits and vegetables, maintaining affordability for all residents, if predicted prices show a rise in the months to come [11].

C. Sharma et al. discuss the problem of agricultural profitability in India. They take into consideration factors like temperature, humidity, pH, rainfall, and other climatic conditions that have an impact on crop yields and, in turn, have an effect on the pricing of fruits, vegetables, and pulses. The authors are aware of the fact that farmers lack information on the choice of crops and anticipated pricing. This issue is addressed by the proposed approach, which predicts crop prices based on previous data trends. Utilizing Decision Tree Regression as a supervised machine learning approach, the method takes into account a number of variables like pH, humidity, precipitation, temperature, and market price [12].

M. Kankar and M.A. Kumar highlight the necessity for technical developments in India's agriculture industry. They propose employing deep learning models like LSTM and BiLSTM as well as deep neural networks like CNN to predict the price of agricultural products. The goal of the project is to reduce market volatility and increase the precision of price prediction for fruits and vegetables by using LSTM models for predicting market prices and a CNN model to classify photos based on variety and quality [13].

R.K. Paul et al. focused on the price prediction of vegetables, particularly Brinjal, at 17 important marketplaces in Odisha, India [14]. The Generalized Neural Network (GRNN), SVR, Random Forest (RF), and Gradient Boosting Machine (GBM) were machine learning methods that were compared with conventional statistical models like the Autoregressive integrated moving average (ARIMA). The findings showed that, when compared to ARIMA, ML approaches, notably GRNN, often displayed higher prediction accuracy. To prove that the ML models outperformed the conventional strategy, the study also used a variety of accuracy metrics and statistical tests. Table I gives details of related literature papers [5]-[14].

TABLE I. Literature Survey

Paper	Research Area	Description
[5] 2019	Time Series	The study focuses on developing mathematical models to analyze price fluctuations in essential agricultural products, specifically mangoes, bananas, cabbage, pechay, and tomatoes in the Philippines' National Capital Region over ten years. These models, including ARIMA and SARIMA, help predict and manage price changes, benefiting consumers, farmers, traders, and policymakers.
[6] 2021	Time Series	The study investigates price transmission dynamics in the Canadian orange and apple markets from 1996 to 2017. It reveals a single, long-term relationship between import and retail prices, with import prices influencing retail prices. Apples exhibit asymmetric price transmission, adjusting faster during constraints than expansions.
[7] 2018	Time Series	The study analyzed ten years of apple pricing and arrivals data from Jammu's Narwal market, revealing a favorable trend with annual price increases of Rs. 220.06 per quintal and peak arrivals from August to January. However, the research did not provide specific price predictions or optimal harvest months for farmers.
[8] 2020	DL	The study compares DL, statistical, and conventional ML models for fresh produce market price prediction. Traditional ML outperforms ARIMA, with CNN-LSTM performing best, predicting prices three weeks ahead.
[9] 2020	ML	The author addresses the complex challenge of forecasting Fresh Produce (FP) pricing, considering factors like limited shelf life, weather impact, and climate change. They aim is to create machine learning models, both traditional and deep learning, using Californian data. A novel aggregated error metric (AGM) is proposed, which blends various evaluation metrics. Stacking ensemble methods, including voting regressor and SVR stacking, are employed to improve predictions.
[10] 2019	ML	The study employed machine learning to classify and predict fruit images with five models. SVM outperformed the others for large and small datasets, highlighting the impact of reduced fruit diversity on algorithm accuracy.
[11] 2019	ML	The author focuses on the rising inflation concern in Bengaluru, Karnataka, India, with a study aiming to assess local fruit and vegetable prices. Seasonal ARIMA models predict future prices to enable proactive price control strategies for affordability.
[12] 2023	ML	The authors examine agricultural profitability in India, considering climate factors (temperature, humidity, pH, rainfall) affecting crop yields and pricing. They address farmers' information gaps by using Decision Tree Regression to predict crop prices based on historical data, including pH, humidity, precipitation, temperature, and market prices.
[13] 2022	DL	The authors underscore the need for technological advancements in India's agriculture sector, advocating the use of LSTM, BiLSTM, and CNN deep learning models to enhance price prediction accuracy, reduce market volatility, and classify fruits and vegetables based on variety and quality in a research project.
[14] 2022	ML/DL	The study focused on predicting Brinjal prices in 17 key markets in Odisha, India, comparing machine learning techniques like GRNN, SVR, RF, and GBM with traditional ARIMA models. Results demonstrated superior prediction accuracy with ML methods, supported by various accuracy metrics and statistical test.

C. Gap Analysis

In the literature studies [5]-[14], various Time Series Analysis, ML and DL, Data Mining and Big Data Analytics, were employed to predict market prices for specific fruits and fresh produce. These studies compared different models, including traditional machine learning, deep learning, and ensemble approaches, to assess their effectiveness in price prediction. The findings indicated that certain models, such as Decision tree regression, SVR and Gradient Boosting, performed well in predicting fruit and fresh produce prices. However, none of the studies specifically addressed the recommendation of the ideal month for farmers to harvest their crops for the highest potential price. This suggests that further research is needed to explore this aspect and provide insights for farmers regarding optimal harvest timing for maximizing profits in the market. While mean absolute error and root mean square error were employed as accuracy measurements in some research, more standardized assessment metrics are required to compare various prediction algorithms. The development of reliable assessment criteria that take predicting accuracy and economic ramifications into account might be a key area of future study.

D. Objective

- Evaluate the effectiveness of several Time Series Analysis, ML and DL regression methods in predicting market prices for certain fruits and vegetables in the Mumbai APMC and Maharashtra, India.
- To identify the most effective regression algorithm among the studied ML and DL techniques for accurately predicting market prices of fruits and vegetables in the specified region.
- Create a predictive model that can accurately estimate prices for fruits and vegetables. This model will help farmers determine the four months with the greatest pricing, allowing them to make smart decisions about when to harvest their crops for maximum profitability in the market.

This paper aims to assist farmers under Mumbai APMC and Maharashtra, India by utilizing and comparing various ML and DL algorithms to determine the optimal harvest timing for maximizing profits. In addition, a standardized metric is employed to compare the performance of all the algorithms. By considering these factors, the research offers valuable guidance for decision-making in the agricultural industry and contributes to maximizing profitability in the market.

III. Experimental Method

The Experimental Method section of this research paper outlines the step-by-step process employed to predict fruit and vegetable prices in the Maharashtra market. To start, an appropriate dataset made up of past price records is gathered. For the purpose of ensuring data integrity, the dataset is then put through a comprehensive cleaning procedure. In order to construct and evaluate models, the dataset is then split into training and test sets. On the training set, a variety of machine learning regression methods are used, and their effectiveness is measured using evaluation metrics like RMSE, MSE, and MAE. These criteria are used to determine the most effective algorithm, which is then used to predict prices. After then, the accuracy of the estimations is checked against current market values. Fig. 1 illustrates a system model at a higher level - Level 0, while Fig. 2 shows an additional representation of system model - low level system model level 1.

The experiment was conducted using Python, a versatile programming language widely used in data science and machine learning.

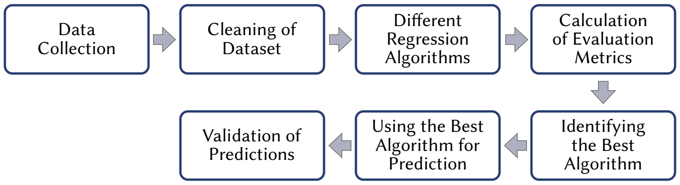

Data Collection
Cleaning of Dataset
Different Regression Algorithms
Calculation of Evaluation Metrics
Identifying the Best Algorithm
Using the Best Algorithm for Prediction
Validation of Predictions

Fig. 1. Level 0 System Model: Higher level.

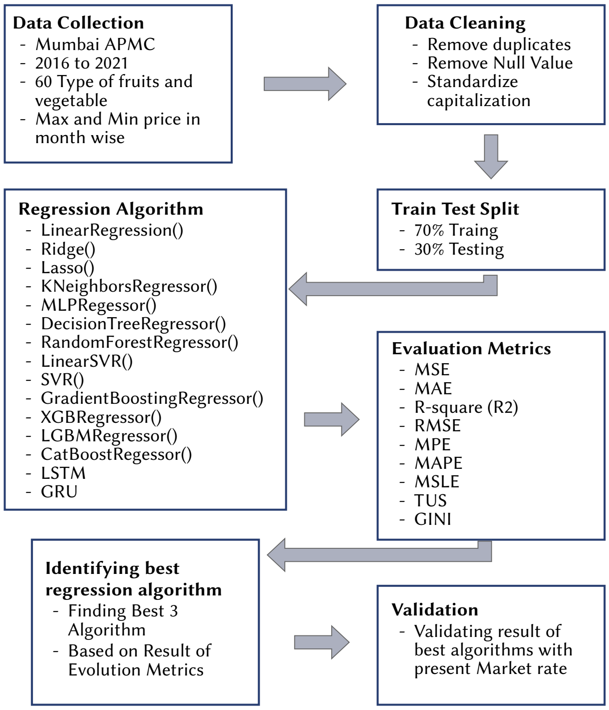


Fig. 2. Level 1 System Model: Low level.

The following Python libraries were utilized:

- **pandas** for data manipulation and analysis.
- **numpy** for numerical operations.
- **sklearn (scikit-learn)** for implementing machine learning models and data preprocessing.
- **xgboost** for the XGBoost model.
- **lightgbm** for the LightGBM model.
- **catboost** for the CatBoost model.

The code was executed on a Jupyter notebook with on Windows 10, 64 bit, NVIDIA GeForce 920M environment. Python version 3.8 was used. All necessary packages were installed via pip or conda package managers.

A. Data Collection

The first step in our study involved collecting relevant data, specifically historical records of fruit and vegetable prices in the Maharashtra market, focusing on a specific time period. We obtained a dataset of fruit and vegetable prices from Mumbai APMC, spanning from April 2016 to March 2021. This dataset includes information on 60 different types of fruits and vegetables.

The collected data consists of the maximum price recorded for each commodity within a given month, as well as the minimum price during the same period. Among the fruits and vegetables included in the dataset are BOR, LIME, GUAVA, KESAR Mangos, PAPAYA, GAJAR (Carate), VANGI (Eggplant), KANDA (Onion), and TAMBATE (Tomato), among others. Table II shows some rows of the datasets.

TABLE II. Dataset

Item	Month	Year	Min Price	Max Price
BOR	April	2016	1600	1900
POMEGRANATE	October	2017	7671	7988
GUAVA	September	2019	3170	4723
TAMBATE	July	2020	3496	4084
PAPAYA	August	2016	968	1432

B. Cleaning of Dataset

After collecting the dataset, we proceeded with a comprehensive cleaning process to address any inconsistencies, errors, or missing values present [15]-[17]. The data cleaning phase involved several tasks, including the removal of duplicate entries, handling missing data through imputation or deletion, and resolving formatting or labeling inconsistencies. This step aimed to ensure the dataset's suitability for subsequent analysis and modeling. To begin, we eliminated any rows containing null values. During non-seasonal periods or months, certain produce items are not available in the market, making their absence understandable. Consequently, removing these null value rows helped maintain data integrity and accuracy. Furthermore, we took measures to address any duplication within the dataset. Duplicate entries can skew analyses and lead to inaccurate results. By identifying and eliminating duplicated data, we ensured that each observation within the dataset was unique and representative. Finally, we devoted attention to resolving inconsistencies in the dataset. This involved rectifying variations in formatting or labeling that could hinder analysis efforts. Overall, through the rigorous data cleaning process, we successfully prepared the dataset for further analysis and modeling by eliminating null values, removing duplicates, and resolving inconsistencies.

We have manually converted the dataset, originally handwritten, to CSV format. Then we used Python programming to clean the data. We used dropna () method of Pandas Data Frame to remove Null data. The Month column of the dataset, which originally included month names as strings, changed into numerical values to enhance computational efficiency. This was done via the replace function in the Pandas library. Regarding feature and target selection, the features (X) used for training include Item, Month, and Year, and the target variables (Y) include Min_Price and Max_Price. Separate target variables Y_min and Y_max were also defined for more specific model training. With the help of scikit-learn's ColumnTransformer and OneHotEncoder, the categorical variable Item was converted into a numerical format that is appropriate for machine learning models through the process of one-hot encoding.

C. Regression Algorithms

We utilized the sklearn library in Python to employ and evaluate various regression algorithms for predicting fruit and vegetable prices in the market. By comparing the results of different algorithms, we aimed to identify the most suitable model for accurate price predictions. Sections 3.3.1 to 3.3.12 present the detailed technical algorithms for the twelve regression techniques referenced in this paper [18]-[21].

1. Linear Regression

The Linear Regression function from sklearn library is a well-liked implementation of this approach in machine learning libraries. Linear regression is a fundamental statistical technique used for predictive modelling [22]-[23]. The provided dataset, which covers the years 2016 through 2021, may be used to estimate fruit and vegetable prices, and linear regression can be a useful technique in this regard. With the use of features like "Fruit or vegetable", "month," and "year," linear regression may determine a connection between these elements and the associated "max price" and "min price" values. Linear regression mathematical relationship between independent and dependent variable is shown in equation (1) [22]-[23], where independent variable are Fruit or vegetable, month, and year and dependent variable are Max price and Min price.

$$y_0 = b_0 + b_1x_1 + b_2x_2 + \cdots + x_nb_n \tag{1}$$

Where: y is still the dependent variable, x_1, x_2,---, x_n are the independent variables, b_o is the intercept term, b_1, b_2, ----, b_n are the coefficients associated with each independent variable. Algorithm 1 gives technical implantation of Linear Regression.

```
Algorithm 1: Linear Regression
1. Initialize weights $w$ and bias $b$ randomly
2. Set learning rate $\alpha$ and number of iterations $N$
3. For $i = 1$ to $N$:
4.   For each training sample $(x, y)$:
5.     Predict $\hat{y} = w * x + b$
6.     Compute error $e = \hat{y} - y$
7.     Update weights: $w = w - \alpha * e * x$
8.     Update bias: $b = b - \alpha * e$
9. Return final weights $w$ and bias $b$
```

2. Ridge Regression (L2 Regularization)

Ridge regression adds a penalty term to the traditional linear regression model as shown in equation (2), forcing the model to not only fit the data but also minimize the sum of squared coefficients [24]. As a result of minimizing the effects of multicollinearity among the features, this regularization strategy stabilizes the model and helps prevent overfitting. Ridge regression can successfully manage the possible strong correlation between months and years, which may affect the price variations of fruits and vegetables. Algorithm (2) shows Ridge Regression (L2 Regularization).

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\,\beta_j\right)^2 - \lambda\sum_{j=1}^{n}\beta_j^2 \tag{2}$$

Where y is dependent variable, x is independent variable, β is coefficient vector to be estimated, λ is the penalty parameter, also known as the tuning parameter, controlling the strength of the penalty term.

```
Algorithm 2: Ridge Regression (L2 Regularization)
1. Input: Feature matrix $X$, target vector $y$, regularization parameter $\lambda$
2. Append a column of ones to $X$ for the intercept term
3. Compute $X$^$T * X$ and $X$^$T * y$
4. Add $\lambda * I$ to $X$^$T * X$  (where $I$ is the identity matrix)
5. Compute the inverse of $(X$^$T * X + \lambda * I)$
6. Compute the ridge coefficients: $\beta = (X$^$T * X + \lambda * I)$^(-1) $* X$^$T * y$
7. Output: Ridge coefficients $\beta$
```

3. Lasso Regression (L1 Regularization)

Lasso regression, on the other hand, adds a penalty term based on the absolute values of coefficients [25] as shown in equation (3). By setting some coefficients to absolutely zero, it accomplishes feature selection, thereby removing less important characteristics. In the context of predicting fruit and vegetable prices, Lasso can pinpoint the characteristics that have the greatest bearing on price changes over time, perhaps emphasizing seasonal trends and particular elements that are key in determining price.

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\,\beta_j\right)^2 - \lambda\sum_{j=1}^{n}\left|\beta_j\right| \tag{3}$$

Where y is dependent variable, x is independent variable, β is coefficient vector to be estimated, λ is the penalty parameter, also known as the tuning parameter, controlling the strength of the penalty term. Algorithm (3) shows Lasso Regression (L1 Regularization).

Algorithm 3: Lasso Regression (L1 Regularization)

```
1. Initialize weights (coefficients) randomly or with zeros.
2. Define a loss function (e.g., Mean Squared Error) and regularization
   strength (lambda).
3. Perform gradient descent or coordinate descent:
   for each iteration {
     a. Compute predictions using current weights.
     b. Compute gradients of the loss function with respect to weights.
     c. Update weights using gradients and regularization term:
        weight = weight - (learning_rate * (gradient + lambda *
        sign(weight)))
   }
4. Repeat until convergence or maximum iterations reached.
```

We can take benefit of the advantages of both techniques by incorporating Ridge and Lasso regularization techniques into linear regression models in order to produce predictions of fruit and vegetable prices that are more reliable and accurate while also reducing the risk of overfitting and dealing with potentially irrelevant features in the dataset.

4. K-Nearest Neighbors (KNN) Regression

Based on historical data, the non-parametric supervised learning technique K-Nearest Neighbours (KNN) regression is used to predict fruit and vegetable prices. KNN Regressor may be used to estimate price trends with a dataset covering the years 2016 through 2021 and include variables such as fruit or vegetable kind, month, year, and maximum and minimum prices. The model estimates prices for certain products at a particular moment by computing the average of the 'k' closest data points with comparable attributes. Because of its simplicity and capacity to identify regional trends in the data, KNN is particularly helpful for predicting the prices of fruits and vegetables, which may experience seasonal or regional variations [26]. The predicted value for a new input x is computed as the average (or weighted average) of the target values of the k nearest neighbors of x and given by equation (4) [26].

$$\hat{y}(x) = \frac{1}{k}\sum_{i=1}^{k} y_i \tag{4}$$

Where x is the input feature vectors (independent variables), y is Target values (dependent variable), k is the number of nearest neighbors, $\hat{y}$ is the predicted target value for the input x. yi is the target value of the ith nearest neighbor of x. Algorithm 4 shows K-Nearest Neighbors (KNN) regression.

Algorithm 4: K-Nearest Neighbors (KNN) regression

```
function KNN_Regression(X_train, y_train, X_test, k):
   for each test sample x_test in X_test:
    Calculate distances d between x_test and all samples in X_train
       Sort distances d in ascending order
       Select the top k samples from X_train based on smallest
       distances
       Predict y_test for x_test as the average of y_train values
       for the selected k samples
   return predicted y_test values
```

5. Multi-Layer Perceptron Regressor

The Multi-Layer Perceptron (MLP) Regressor is a class of neural network used for regression problems, we use MLP Regressor function from sklearn. neural network library to use MLP [24]. This model may be used to project future price patterns based on previous data in the context of projecting fruit and vegetable prices. The 2016–2021 dataset, which includes characteristics like fruit or vegetable kind, month, year, maximum price, and minimum price, enables the MLPRegressor to learn intricate patterns and correlations within the data and provide price prediction. By addressing non-linear correlations between input characteristics and target pricing effectively, this strategy enables companies to make well-informed decisions, optimize their supply chains, and anticipate market changes. For a single-hidden-layer MLP regressor with h neurons in the hidden layer, the output of the hidden layer can be calculated as shown in equation (5) [27].

$$Z = \emptyset(XW^{(1)} + b^{(1)}) \tag{5}$$

Where $W^{(1)}$ is the weight matrix connecting the input layer to the hidden layer, with dimensions (n, h), $b^{(1)}$ is the bias vector for the hidden layer, with dimensions $(1, h)$, ϕ is the activation function applied element-wise to the weighted sum.

The output of the MLP Regressor can then be calculated as shown in equation (6).

$$\hat{Y} = ZW^{(2)} + b^{(2)} \tag{6}$$

Where, $W^{(2)}$ is the weight matrix connecting the hidden layer to the output layer, with dimensions $(h, 1)$, $b^{(2)}$ is the bias for the output layer. Algorithm 5 shows Multi-Layer Perceptron Regressor.

Algorithm 5: Multi-Layer Perceptron Regressor

```
1. Initialize weights randomly
2. Define activation function (e.g., sigmoid, ReLU)
3. Define learning rate (alpha), number of layers (L), number of
   neurons per layer (N)
4. For each epoch:
    For each training example (X, y):
      Forward pass:
        Calculate output of each neuron in each layer using current
        weights
         Apply activation function to each neuron's output
      Backward pass:
         Calculate error derivative with respect to output
         Update weights using gradient descent
5. Repeat step 4 until convergence or maximum number of epochs
   reached
6. Output the trained MLP model
```

6. Decision Tree Regression

The Decision Tree Regressor is a regression technique that uses historical data from 2016 to 2021 and a Decision Tree algorithm to predict fruit and vegetable prices. The dataset is recursively partitioned using variables like type (fruit or vegetable), month, and year to create a tree-like model. Price, as indicated by maximum and minimum prices, is the dependent variable to be predicted. The Decision Tree can manage non-linear patterns in price swings because it can record complicated correlations between characteristics and the target [28]-[29]. The mathematical equation for decision tree regression is expressed in equation (7).

$$\hat{y} = \sum_{i=1}^{N} w_i \,.\, I(x \in R_i) \tag{7}$$

Where, $\hat{y}$ represents the predicted output. N is the number of leaf nodes, W_i is the predicted value at leaf node, R_i denotes the region defined by the ith leaf node. $I(x \in R_i)$ is an indicator function that returns 1 if the input x belongs to the region R_i, and 0 otherwise. Algorithm 6 shows Decision Tree Regression.

Algorithm 6: Decision Tree Regression

```
1.  DecisionTreeRegression(data, max_depth, min_samples_split):
2.      if max_depth == 0 or len(data) < min_samples_split:
3.          return Leaf Node (mean(data. target))
4.      else:
5.          best_split = find_best_split(data)
6.          if best_split == None:
7.              return LeafNode(mean(data.target))
8.          left_data, right_data = split(data, best_split)
9.          left_subtree = Decision Tree Regression (left_data, max_depth - 1, min_samples_split)
10.         right_subtree = Decision Tree Regression (right_data, max_depth - 1, min_samples_split)
11.         return Decision Node (best_split, left subtree, right_subtree)
12. find_best_split(data):
13.     best_split = None
14.     best_mse = infinity
15.     for each feature in data.features:
16.         for each threshold in unique(data[feature]):
17.             left_data, right_data = split(data, (feature, threshold))
18.             mse = weighted_mse(left_data, right_data)
19.             if mse < best_mse:
20.                 best_mse = mse
21.                 best_split = (feature, threshold)
22.     return best_split
```

In algorithm 6 Lines 1-11 the main recursive function for building the decision tree is defined. It stops if the maximum depth is reached or the number of samples is too small. Otherwise, it finds the best split and recursively builds left and right subtrees. Lines 12-22 define the function to find the best split based on the minimum mean squared error (MSE). This function evaluates all possible splits and returns the one with the lowest MSE.

7. Random Forest Regression

Random Forest Regressor is a popular ML algorithm based on the Random Forest ensemble method. It works for regression jobs and has the ability to handle both categorical and numerical data. It becomes a useful tool for solving a variety of regression issues by integrating numerous decision trees, which decreases overfitting and increases prediction accuracy [30]. The model can efficiently analyze features like the type of fruit or vegetable, month, year, maximum price, and minimum price. The model can predict future price changes by using historical information, which helps producers, suppliers, and customers make wise decisions and adjust to market changes. The mathematical equation for Random Forest Regression can be summarized as given in equation (8).

$$\hat{y} = \frac{1}{2}\sum_{i=1}^{N} f(x_i) \tag{8}$$

Where $\hat{y}$ is the predicted output, N is the number of trees in the forest, $f(x_i)$ represents the output (prediction) of the ith decision tree in the forest for input x. Algorithm 7 shows Random Forest Regression.

Algorithm 7: Random Forest Regression

```
Input: Training data (X_train, y_train), number of trees N, max depth D, sample size S
Initialize an empty list of trees: forest = [ ]
for i = 1 to N do
    Sample with replacement S data points from (X_train, y_train)
    Build a Decision Tree Regression T on the sampled data with max depth D
    Add tree T to the forest
end for

Function Predict(X_test):
    Initialize predictions = [ ]
    for each tree T in forest do
        pred = T.predict(X_test)
        Add pred to predictions
    end for
    return average(predictions)
end Function
```

8. Support Vector Machine: Linear Kernel and Radial Basis Function Kernel

Multidimensional data may be handled by Support Vector Machines (SVM) using linear and RBF (Radial Basis Function) kernels, making them appropriate for analyzing features like fruit or vegetable, month, year, maximum price, and minimum price. In order to approximate linear correlations between features and prices, LinearSVR function from sklearn.svm library uses a linear kernel that performs well for datasets with linearly separable classes. However, SVR with an RBF kernel may detect non-linear patterns in the data, which is crucial for detecting intricate interactions and seasonal changes that may have an impact on the pricing of fruits and vegetables [31]. The decision function for the linear kernel SVM is given by equation (9).

$$f(x) = sign(w.x + b) \tag{9}$$

Where w is the weight vector, x is the input vector, b is the bias term, sign(·) is the sign function, returning -1 for negative values and 1 for non-negative values. The decision function for the RBF kernel SVM is given by equation (10).

$$f(x) = sign(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b) \tag{10}$$

Where α_i are the Lagrange multipliers obtained during training, y_i are the class labels of the training data, x_i are the support vectors, $K(x_i, x)$ is the kernel function, which in the case of RBF kernel is given by equation (11).

$$K(x_i, x) = exp(-\gamma \|x_i - x\|^2) \tag{11}$$

Where γ is the kernel parameter and ‖·‖ denotes the Euclidean distance. Algorithms 8 and 9 show SVM Liner Kernel and SVM Radial Basis function.

Algorithm 8: Support Vector Machine: Linear Kernel

```
Input: Training data $(X, y)$, regularization parameter $C$
Initialize: weights $w$, bias $b$
while not converged:
    for each $(xi, yi)$ in $(X, y)$:
        if $yi * (w \cdot xi + b) < 1$:
            $w = w + \eta * (yi * xi - 2 * \lambda * w)$
            $b = b + \eta * yi$
        else:
            $w = w - \eta * 2 * \lambda * w$
Output: weights $w$, bias $b$
```

Algorithm 9: Support Vector Machine: Radial Basis Function Kernel

```
Input: Training data $(X, y)$, regularization parameter $C$, kernel parameter $\gamma$
Initialize: Lagrange multipliers $\alpha$, bias $b$
while not converged:
    for each $(xi, yi)$ in $(X, y)$:
        compute kernel: $K(xi, xj) = exp(-\gamma * ||xi - xj||\hat{}2)$
        if $yi * (\Sigma \alpha j\, yj\, K(xj, xi) + b) < 1$:
            $\alpha i = \alpha i + \eta * (1 - yi * (\Sigma \alpha j\, yj\, K(xj, xi) + b))$
            $b = b + \eta * yi$
        else:
            $\alpha i = \alpha i$
Output: Lagrange multipliers $\alpha$, bias $b$
```

9. Gradient Boosting

Gradient Boosting Regressor is a popular machine learning algorithm used for regression tasks. It systematically builds several weak learners (often decision trees), with each tree attempting to fix the flaws of the one before it [32]. Gradient Boosting can successfully handle complicated interactions between data like month, year, max price, and min price in the context of fruit and vegetable price prediction. It can predict for various fruits and vegetables across the 2016–2021 dataset by capturing non-linear patterns. The mathematical equation for Gradient Boosting Regressor can be represented by equation (12).

$$F(x) = \sum_{m=1}^{M} \gamma_m h_m(x) \qquad (12)$$

Where $F(x)$ is the final prediction function, M is the number of weak learners (trees) in the ensemble, $h_m(x)$ represents the output of the m^{th} weak learner, typically a decision tree, γ_m is the weight (or learning rate) assigned to the m^{th} weak learner. Algorithm 10 shows gradient boosting Regressor.

Algorithm 10: Gradient Boosting Regressor

```
Initialize $F\_0(x) = y_{mean}$
For $m = 1$ to $M$ do:
    Compute residuals $r\_m = y - F_{(m\text{-}1)}(x)$
    Fit a base learner $h_{m(x)}$ to residuals $r_m$
    Compute optimal step size $\gamma\_m$
    Update model: $F_{m(x)} = F_{(m\text{-}1)}(x) + \gamma_m\, h_m(x)$
End For
Output final model $F_{M(x)}$
```

10. XGBoost

The XGBRegressor is an optimized implementation of Gradient Boosting. Due to its regularization methods, parallel processing, and the handling of missing data, it delivers improved performance and efficiency [33]. XGBoost can use the dataset's temporal properties (month and year) to capture the impacts of seasonality and provide accurate price estimates for fruit and vegetable prices. The mathematical equation for predicting the target variable y using XGBRegressor is shown in equation (13).

$$\hat{y_\iota} = \emptyset(x_i) = \sum_{k=1}^{k} f_k(x_i), \qquad f_k \in \mathrm{F} \qquad (13)$$

Where $\hat{y}_i$ is the predicted value for the i^{th} instance, x_i represents the features of the i^{th} instance, K is the number of trees in the ensemble, f_k is an individual decision tree from the ensemble. F is the space of all possible decision trees. Algorithm 11 shows XGBoost regressor

Algorithm 11: XGBoost Regressor

```
Initialize model with a constant value: $\emptyset(x_i) = \lambda$
For each iteration $t = 1$ to $T$:
    Compute the negative gradient of the loss function for each data point: $g\_i = \partial L(y_i, \emptyset(x_i))/\partial\emptyset(x_i)$
    Fit a regression tree to the negative gradients: $T_t = TreeFit(\{(x_i, g_i)\})$
    Update the model: $\emptyset(xi) = \emptyset(xi) + \eta * Tt(xi)$
End for
```

11. Light GBM

LGBMRegressor is another gradient boosting algorithm known for its high speed and memory efficiency. By splitting trees into leaf-wise rather than levels, it uses less processing power [34]. LightGBM is a useful model for estimating price trends because it can effectively manage the temporal features of the dataset and accommodate fluctuations in pricing for various produce over time in the context of fruit and vegetable price prediction. LGBMRegressor Mathematical equation is given by equation (14).

$$\mathrm{Y} = \mathrm{BaseTree(x)} - \mathrm{lr.Tree_1(x)} - \mathrm{lr.Tree_2(x)} - \mathrm{lr.Tree_3(x)} \qquad (14)$$

Where Y represents the predicted target value, Base Tree(x) is the output of the base tree, which is essentially the initial prediction made by the model. $\mathrm{Tree_1(x)}$, $\mathrm{Tree_2(x)}$, $\mathrm{Tree_3(x)}$ are the contributions from individual trees. lr denotes learning rate. Algorithm 12 shows Light Gradient Boosting Model Regressor.

Algorithm 12: LGBMRegressor

```
Initialize dataset $D$, number of trees $T$, learning rate $lr$
Initialize model: $Y = BaseTree(x)$
For $t = 1$ to $T$:
    Compute residuals $r = Y - lr * Tree_t(x)$
    Fit regression tree to residuals: $Tree_t(x)$
    Update model: $Y = Y - lr * Tree_t(x)$
End For
Output: Final model Y
```

12. CatBoost

CatBoost Regressor is a popular algorithm for regression tasks, designed to handle both numerical and categorical features [34]. The fruit and vegetable price prediction dataset, which contains categorical variables like month, is a good fit for it since it can handle categorical data without the requirement for explicit encoding. The model is effective for large-scale prediction tasks when the verbose option is set to 0, which guarantees that it operates silently and prevents unnecessary output during training and prediction. It is given by equation (15) and algorithm 13 .

$$L(t,a) = \frac{\sum_{i=0}^{N} w_i |a_i - t_i|}{\sum_{i=0}^{N} w_i} \tag{15}$$

Where, $L(t, a)$ represents the loss function. t_i is the true target value for the i^{th} sample, a_i is the predicted value for the i^{th} sample, w_i denotes the weight assigned to the i^{th} sample (usually equal to 1).

The range of regression algorithms discussed above offers a diverse set of tools for predicting fruit and vegetable prices based on historical data spanning 2016 to 2021. Fundamental methods for modelling variations in prices include linear regression, ridge, and lasso, while K-Nearest Neighbours captures regional and seasonal patterns. MLPRegressor and other neural network techniques uncover complex patterns in the data. Support Vector Machines handle multidimensional data and linear/RBF patterns, whereas Decision Tree and Random Forest handle non-linear connections. Gradient Boosting, XGBoost, LightGBM, and CatBoost optimize performance, scalability, and efficiency, making them suitable for large datasets.

Algorithm 13: CatBoost Regressor

```
1. Initialize ensemble of decision trees
2. For each tree:
3.    Initialize leaves with average target value
4.    For each feature:
5.        For each split point:
6.            Calculate loss reduction using **equation 15**
7.    Choose the best split based on loss reduction
8.    Update leaf values based on targets within each leaf
9. Train until convergence or maximum number of iterations
10. Output ensemble of decision trees
```

D. Evaluation Metrics for Regression Algorithms

Regression algorithms use evaluation metrics to measure the effectiveness and precision of the model's predictions in relation to the actual target values. Mean Absolute Error (MAE) [35]-[36], Mean Squared Error (MSE) [35]-[36], and R-squared (R2) or Coefficient of Determination [36] are the three most often used assessment metrics for regression models employed in related studies [9]-[34]. In this paper we have also compared different regression algorithm with Root Mean Square Error (RMSE) [35]-[36], Mean Percentage Error (MPE) [36], Mean Absolute Percentage Error (MAPE) [36], Huber Loss (HL) [37], Mean Squared Logarithmic Error (MSLE) [38], Theil's U Statistic (TUS) [39], Gini Coefficient (Gini) [40]. The evaluation of each regression model was carried out with the help of the evaluation metrics mentioned above, and we utilized numpy and the sklearn. metrics package in Python.

1. Mean Absolute Error (MAE)

MAE calculates the average absolute difference between the predicted values and the actual target values. It measures the average magnitude of errors without considering their direction [35]. Equation (16) gives MAE.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{16}$$

Where, n is the number of data points, y_i is the actual target value for data point i, $\hat{y}_i$ is the predicted value for data point i.

2. Mean Squared Error (MSE)

The average of the squared differences between the predicted values and the actual target values is computed using MSE. It is frequently applied in different regression techniques and penalizes larger errors more severely than MAE [35]. Equation (17) gives MSE.

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{17}$$

Where n is the number of data points, y_i is the actual target value for data point i, $\hat{y}_i$ is the predicted value for data point i.

3. R-squared (R2) or Coefficient of Determination

R-squared calculates the proportion of the target's variance that can be predicted from the model's independent variables (features). The value ranges from 0 to 1, with 0 denoting that the model explains no variation and 1 denoting a perfect match [35]. R-squared is computed using equation (18).

$$R^2 = 1 - \frac{SS_{residuals}}{SS_{total}} \tag{18}$$

Where $SS_{residuals}$ is the sum of squared residuals (difference between actual and predicted values), SS_{total} is the total sum of squares (variance of the actual target values).

4. Root Mean Square Error (RMSE)

RMSE quantifies the average discrepancy between the projected and actual prices. The calculation involves finding the square root of the average of the squared discrepancies between the actual (y_i) and predicted ($\hat{y}_i$) prices as shown in equation (19).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{19}$$

Here 'n' represents the number of predictions. A lower RMSE indicates higher prediction accuracy.

5. Mean Percentage Error (MPE)

MPE is a metric employed to assess the precision of predictions. The function computes the mean percentage deviation between projected and real prices. The equation (20) represents MPE.

$$MPE = \frac{1}{n}\sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)}{y_i} * 100 \tag{20}$$

Where n is the number of data points, y_i is the actual target value for data point i, $\hat{y}_i$ is the predicted value for data point i.

6. Mean Absolute Percentage Error (MAPE)

MAPE is a metric that quantifies the accuracy of predictive models. The algorithm computes the mean absolute percentage deviation between predicted and observed prices. The equation (21) shows MAPE.

$$MPE = \frac{1}{n}\sum_{i=1}^{n} \left|\frac{(y_i - \hat{y}_i)}{y_i}\right| * 100 \tag{21}$$

Where n is the number of data points, y_i is the actual target value for data point i, $\hat{y}_i$ is the predicted value for data point i. MAPE expresses error as a percentage, which simplifies its interpretation. A smaller MAPE signifies a higher level of accuracy in predictions.

7. Mean Squared Logarithmic Error (MSLE)

The MSLE quantifies the average squared discrepancy between the natural logarithm of the predicted values and the actual values. It is commonly used in fruit prediction of prices, where the estimates can vary greatly in magnitude. MSLE is computed using equation (22).

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (log(y_i + 1) - log(\hat{y}_i + 1))^2 \tag{22}$$

Where n is the number of data points, y_i is the actual target value for data point *i*, $\hat{y}_i$ is the predicted value for data point *i*. This measure imposes equal fines for both underestimation and overestimation, rendering it appropriate for datasets that exhibit skewness. A lower MSLE value suggests more accuracy in forecasting fruit prices.

8. Loss (HL)

The Huber loss function, commonly used in regression applications such as fruit price prediction, combines the resilience of mean squared error (MSE) with the responsiveness of mean absolute error (MAE). It reduces the influence of extreme values on the model's performance. Equation (23) shows HL and is given by $L_\delta(y - \hat{y})$.

$$L_\delta(y - \hat{y}) = \begin{cases} \frac{1}{2}(y - \hat{y})^2, & |y - \hat{y}| \le \delta \\ \delta(|y - \hat{y}| - \frac{1}{2}\delta), & Otherwise \end{cases} \tag{23}$$

Where y is actual target value, $\hat{y}$ is the predicted value, δ is a parameter that determines the threshold beyond which the loss becomes linear rather than quadratic. We have use default value δ as 1. The Huber loss function offers a well-balanced method by penalizing significant errors in a linear manner within a specified tolerance (δ), and in a quadratic manner beyond that tolerance. This guarantees enhanced robustness against outliers values while also achieving efficient model optimization.

9. Theil's U Statistic (TUS)

Theil's U Statistic (TUS) is a metric used in econometrics to evaluate the accuracy of predictions, specifically in the context of predicting fruit prices. The process involves comparing the observed values with the expected values, taking into account both bias and variability. TUS equation is shown in equation (24).

$$TUS = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)}{\sum_{i=1}^{n} y_i}} \tag{24}$$

Where, *n* is the number of data points, y_i is the actual target value for data point *i*, $\hat{y}_i$ is the predicted value for data point *i*. A TUS number near to 0 suggests accurate predictions, whereas higher levels indicate more prediction errors.

10. Gini Coefficient (Gini)

The Gini Coefficient (Gini) is a statistical indicator that measures the level of inequality in a given distribution. It is commonly employed in the field of economics for predicting fruit prices. The scale spans from 0, indicating complete equality, to 1, representing extreme disparity. Equation (25) shows Gini Coefficient.

$$Gini = \frac{\sum_{i=1}^{n}(2i - n - 1)(y_i - \hat{y}_i)}{\sum_{i=1}^{n} y_i} \tag{25}$$

Where *n* is the number of data points, y_i is the actual target value for data point i, $\hat{y}_i$ is the predicted value for data point *i*. A higher Gini coefficient indicates a higher level of inequality, suggesting the presence of possible price differences in the fruit market.

To obtain a comparison evaluation of the regression model's performance, it is crucial to combine these measures. R-squared provide information on how effectively the model accounts for the variance in the target variable, whereas MAE and MSE give a sense of the absolute error size. When working with complicated models and huge datasets, R-squared must be utilized with caution because it may occasionally be deceptive and should be combined with additional metrics [35]-[36].

E. Identifying and Validating Best Regression Algorithm

After evaluating and comparing various regression algorithms, we will select the most suitable one for predicting fruit and vegetable prices. By analyzing the data, we will identify the top 4 months with the highest prices for specific produce. This valuable information will empower farmers to plan their crop harvesting and cultivation strategically, ensuring they obtain the best possible prices for their products. Additionally, policymakers can benefit from this prediction to make informed decisions and support the agricultural sector. The accuracy of this prediction will be validated against the current market price of tomatoes, ensuring its reliability and usefulness in real-world scenarios.

Next section shows basic experimentation and data analysis results which include analysis of price of tomato and mosambi, Regression Algorithms and prediction of price from best regression algorithms.

IV. Result and Discussion

In this section, we present the predicted prices of fruits and vegetables in the Maharashtra market for the next few year based on our research and analysis. We obtained a dataset of fruit and vegetable prices from Vasai Market, spanning from April 2016 to March 2021. This dataset includes information on 60 different types of fruits and vegetables. Detail about dataset is given in section 3. The predictions aim to provide insights into potential price fluctuations and aid market stakeholders in making informed decisions.

A. Analysis of Price of Tomato and Mosambi From April 2016 to March 2021

In this section, we examine historical pricing patterns for the previous few years for tomatoes and mosambi (sweet lime). The results are based on the dataset discussed in section 3, which shows the rise in particular months and fall in particular months in prices for tomatoes and mosambi.

Fig. 3 shows historical tomato pricing data over the last several years. Different time periods are shown by the x-axis, while matching prices in the Mumbai APMC are shown on the y-axis. Our examination of the data shows a yearly increase tendency in tomato prices in the months of June, July, and August. This increase can be linked to a number of things, including changes in the dynamics of supply and demand, climatic variables that impact growing, the cost of transportation, and competition.

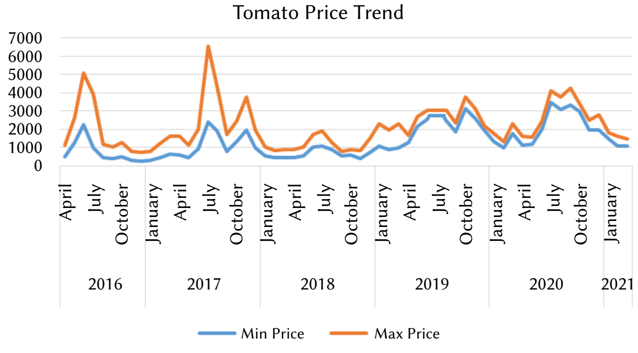


Fig. 3. Tomato Price Trend.

Fig. 4 shows the historical price changes for Mosambi during the given time period. The graph depicts a rising tendency in Mosambi pricing, similar to the pattern in tomato prices. Mosambi prices might rise as a result of market demand, supply chain interruptions, shifting customer tastes, and changes in farming practices. Additionally, outside variables like climatic changes and the dynamics of the worldwide market might have influenced price developments. Consumers, companies, and agricultural stakeholders may be impacted by the steady increase in Mosambi prices. It could have an impact on consumer choices, dynamics of export and import, and fruit growers' financial success.

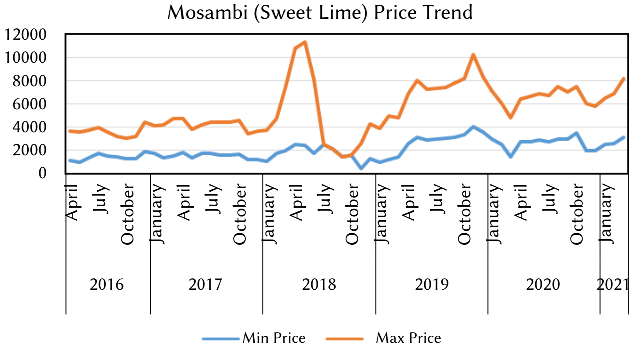


Fig. 4. Mosambi (Sweet Lime) Price Trend.

B. Regression Algorithms

In Section 3 of the research, the dataset preprocessing and regression analysis pipeline were comprehensively detailed. The dataset consists of 60 distinct fruits and vegetables collected over a 5-year period, resulting in a total of 3,600 data entries, with each year contributing 12 rows of data per fruit/vegetable type. First, all null values were removed, resulting in a refined dataset containing 3,115 entries. Following this, the categorical variable "Month" was transformed. Originally represented by the names of the months (e.g., January, February), this variable was converted into numerical ranging from 1 to 12. To handle categorical features effectively during the regression analysis, the one-hot encoding technique was applied. This transformation converted categorical features into binary vectors, making them compatible with various regression algorithms [41]. The dataset was then divided into training and testing sets. Specifically, 70% of the data (2,180 entries) was allocated for training purposes, while the remaining 30% (935 entries) was reserved for testing the trained models. This separation ensures that the models are evaluated on unseen data, providing a more accurate assessment of their generalization performance. To further enhance the robustness of the regression models, k-fold cross-validation was employed during the training process. This technique involves dividing the training dataset into 'k' folds or subsets. The model is then trained 'k' times, each time using a different fold as the testing set and the remaining fold for training. The final performance metric is calculated by averaging the results from each iteration. In this research, we chose a value of k =5. Additionally, to ensure uniformity in the scale of input features, the dataset underwent standardization using the Standard Scaler [42]. This preprocessing step is particularly essential for regression algorithms that are sensitive to variations in the scale of input features, promoting stable and reliable model training across different features. By incorporating k-fold cross-validation and standardization into the training process, the research aims to provide a more robust evaluation of the regression models' performance, accounting for potential variations and improving their generalization capabilities.

All 13 listed regression algorithms were employed, and the training data was used as input for each of them. We have also added 2 more DL algorithm LSTM (Long Short-Term Memory) and GRU (Gate Recurrent Unit). The performance of the models was evaluated using three key metrics: MSE, MAE, and R-square (R^2). These metrics provide insights into the accuracy and goodness-of-fit of the regression models. Table III and Table IV presented the performance metrics of all the algorithms, providing a comprehensive comparison of their predictive capabilities. The results shed light on the strengths and weaknesses of each algorithm in capturing the underlying patterns and relationships in the dataset.

TABLE III. Performance Metrics of All the Algorithms (MSE, MAE, R-Square, RMSE, MPE)

Model	MSE	MAE	R-square (R^2)	RMSE	MPE
LinearRegression()	33.97	44.54	0.56	5.83	-17.89
Ridge()	34.03	44.53	0.56	5.83	-19.84
Lasso()	34.00	44.58	0.56	5.83	19.54
KNeighborsRegressor()	20.81	34.20	0.73	4.56	-21.76
MLPRegressor()	80.62	85.55	0.00	8.98	-40.65
DecisionTreeRegressor()	0.00	0.00	1.00	0.00	0.00
RandomForestRegressor()	3.94	11.53	0.95	1.99	-3.28
LinearSVR()	94.28	100.0	0.00	9.71	-24.48
SVR()	100.0	90.10	NA	10.00	-50.13
GradientBoostingRegressor()	16.55	49.86	0.82	4.07	-39.98
XGBRegressor()	2.67	18.52	0.97	1.63	-7.18
LGBMRegressor()	31.06	45.05	0.67	5.57	-18.91
CatBoostRegressor()	5.17	26.92	NA	2.27	-12.39
LSTM	60.24	20.48	NA	7.76	-18.21
GRU	55.54	20.08	NA	7.45	-17.82

TABLE IV. Performance Metrics of All the Algorithms (MAPE, HL, MSLE, TUS, GINI)

Model	MAPE	HL	MSLE	TUS	GINI
Linear Regression ()	39.23	51.67	0.27	0.31	0.39
Ridge()	39.86	51.67	0.27	0.32	0.39
Lasso()	39.78	51.72	0.26	0.32	0.38
K Neighbors Regressor ()	35.53	39.68	0.18	0.28	0.41
MLPRegressor()	52.89	100.00	0.83	0.98	0.31
Decision Tree Regressor ()	0.00	0.00	0.00	0.00	0.04
Random Forest Regressor ()	8.35	13.05	0.02	0.06	0.04
Linear SVR ()	74.83	82.52	0.67	0.55	0.45
SVR()	88.80	80.29	0.65	0.70	0.76
Gradient Boosting Regressor ()	56.65	47.05	0.28	0.45	0.38
XGBRegressor()	17.38	17.52	0.07	0.14	0.04
LGBMRegressor()	35.94	44.78	0.18	0.29	0.39
CatBoost Regressor ()	26.60	26.63	0.11	0.21	0.42
LSTM	64.26	87.26	0.52	0.47	0.59
GRU	63.85	86.13	0.51	0.43	0.57

For simplicity of comparison, the MAE values have been scaled down to a range of 0 to 100 and the MSE values to a range of 0 to 1000. Fig. 5 shows comparison of the different algorithms, all metrics are scaled down to 0 to 1 in the comparison chart.

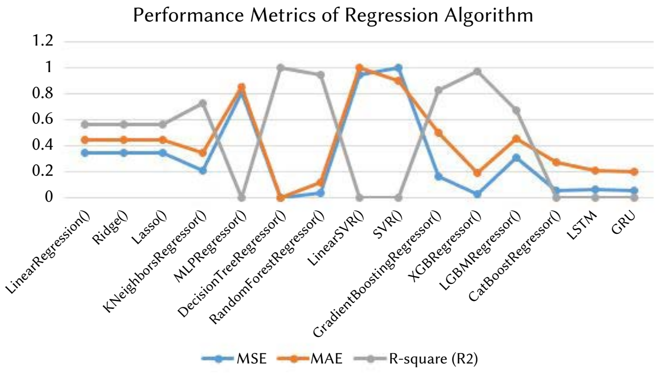


Fig. 5. Performance Metrics of regression Algorithm.

1. Linear Regression (), Ridge (), and Lasso ()

The MSE, MAE, and R2 values obtained from these three linear regression methods were comparable, demonstrating equivalent performance on the dataset. The models' estimated R2 value of 0.564 indicates that they account for around 56.4% of the variation in the target variable. This indicates a moderate level of predictive power, implying that the models capture a substantial portion of the variability present in the data.

2. KNeighbors Regressor ()

With reduced MSE and MAE and a higher R2 value of 0.728, the KNeighbors Regressor beat the linear regression models, indicating that this algorithm provides a better fit to the data, potentially due to its ability to capture nonlinear relationships and complex patterns.

3. MLP Regressor ()

Dataset was divided in batch size of 64 and total iteration completed by algorithm was 100. In comparison to other models, the MLPRegressor displayed much higher MSE and MAE values, indicating that it underperformed on this dataset. Additionally, the target variable's variation is only partially explained by this model, as seen by the low R2 value of 0.001.

4. Decision Tree Regressor ()

The Decision Tree Regressor performed remarkably well, achieving an MSE and MAE of 0.000, which is likely due to overfitting. The perfect R2 value of 1.000 indicates that this model perfectly fits the data, but it might not generalize well to new data. Let's denote the true target values as y_i and the predicted values by the decision tree as $\hat{y}_i$. The total sum of squares (TSS) represents the total variance in the target variable. TSS is given by equation (26).

$$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{26}$$

Where $\bar{y}$ is the mean of the true target values. The residual sum of squares (RSS) measures the unexplained variance by the model. RSS is given by equation (27).

$$RSS = \sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{27}$$

The R-squared value is given by equation (28).

$$R^2 = 1 - \frac{RSS}{TSS} \tag{28}$$

For a perfect fit, (R^2= 1), indicating that the model explains all the variance. Since the Decision Tree Regressor () has an R^2 value of 1.00, it means that it perfectly fits the training data.

5. Random Forest Regressor ()

The Random Forest Regressor produced low MSE and MAE values, indicating good performance. The relatively high R2 value of 0.947 suggests that this model explains a significant portion of the variance in the target variable.

6. Linear SVR () and SVR ()

Both Support Vector Regressor algorithms showed relatively high MSE and MAE values, with very low R2 values. The R2 values, which measure the proportion of variance in the dependent variable that is predictable from the independent variables, were very low. These findings suggest that the SVR models struggled to capture the underlying patterns in the dataset effectively, leading to subpar regression performance.

7. Gradient Boosting Regressor ()

The Gradient Boosting Regressor achieved a reasonable MSE and MAE, along with a relatively high R2 value of 0.825, indicating good performance and a relatively better fit compared to some other models. This performance underscores the effectiveness of Gradient Boosting techniques in handling complex relationships within the data and making accurate predictions.

8. XGB Regressor ()

The XGBRegressor showed a low MSE and MAE, as well as a high R2 value of 0.972, indicating excellent performance and a strong ability to explain variance in the target variable. This level of performance indicates that the XGBRegressor successfully captures complex relationships within the data, providing reliable predictions.

9. LGBM Regressor ()

The LGBMRegressor had a moderate MSE and MAE, along with an R2 value of 0.671, suggesting a reasonable fit to the data. Overall, these results imply that the LGBMRegressor is capturing a significant portion of the underlying patterns in the data.

10. CatBoost Regressor ()

The CatBoost Regressor performed quite well, with relatively low MSE and MAE values, indicating good performance on the dataset. The model's adeptness in capturing intricate relationships within the data translates to accurate predictions. Its boosted decision tree architecture contributes to its prowess by iteratively improving upon weaknesses, enhancing predictive accuracy with each iteration. Overall, the CatBoost Regressor emerges as a formidable choice for tasks requiring precise regression, showcasing its reliability and effectiveness in real-world applications.

11. LSTM

The dataset was divided in batch size of 32 and total iteration completed by algorithm was 50. The LSTM model achieved a mean squared error (MSE) of 60.235 and a mean absolute error (MAE) of 20.480. These relatively low error values suggest that the LSTM model provides reasonably accurate predictions.

12. GRU

The dataset was divided in batch size of 32 and total iteration completed by algorithm was 50. The GRU model also performed well with an MSE of 55.54 and an MAE of 20.081. The low MSE and MAE values indicate that the GRU model is capable of making relatively accurate predictions.

Based on the performance metrics, the Decision Tree Regressor () stands out as the top-performing model, with an MSE and MAE of 0.000, indicating a perfect fit to the data. While this might seem impressive at first glance, it is essential to keep in mind that such a result could be a sign of overfitting, where the model memorizes the training data but fails to generalize to new, unseen data.

The Random Forest Regressor () and XGB Regressor are strong contenders for a more balanced selection. Both models exhibited quite low MSE and MAE values, indicating strong performance and some degree of prediction accuracy. The strong R2 values demonstrated their ability to fit the data accurately by explaining a substantial amount of the variance in the target variable.

When examined more closely, the XGBRegressor emerges as a highly advantageous choice. It demonstrated the best accuracy and precision in predicting the target variables (max price and min price) with the lowest MSE and MAE values after Decision Tree Regressor among all other models. Furthermore, the XGBRegressor proves to be a solid option for this regression task, evident from its high R2 value of 0.972, indicating that it can explain a sizable percentage of the variance in the target variable.

Considering the overall performance and the balance between accuracy and generalization, the XGBRegressor appears to be the most appropriate regression method for this particular dataset and research challenge. It stands out as the recommended option due to its capacity to precisely anticipate the target variables while still maintaining good generalization.

Overall, the findings presented valuable insights for understanding the predictive power of various models, facilitating better decision-making and potential applications in real-world scenarios.

C. Predicting Values From Best Regression Algorithm

In this study, we employed three top-performing algorithms, XGBRegressor, Random Forest, and Decision Tree, to predict tomato prices for the year 2023. Our proposed method effectively identified the four most lucrative months in terms of market prices. This information can greatly assist farmers in making informed decisions about the optimal timing for tomato harvesting. The results presented in Table V highlight the months with the highest price potential, empowering farmers to maximize their profits and optimize crop cultivation strategies.

TABLE V. Tomato Price Prediction

Sr. No.	XGBRegressor		Random Forest		Decision Tree	
	Month	Price	Month	Price	Month	Price
1	9	3138.57	9	3906.23	9	4258
2	7	3133.95	8	3647.96	7	4084
3	10	3131.60	7	3560.89	8	3744
4	8	2942.82	10	3488.16	10	3400

Result in Table V shows that all three algorithm predicted the same highest four months in which farmer can get more profit. This means that prediction is good by comparing best three regression algorithm. This implies that our prediction is robust. Additionally, Table VI shows the average of the predicted prices for these four months, providing valuable insights for farmers to capitalize on potentially more profitable periods.

TABLE VI. Predicted Price Trends: Average of Top 3 Algorithm

Month	Average
7	3592.95
8	3444.93
9	3767.60
10	3339.92

To validate the prediction that tomatoes became expensive in India from the month of July, we refer to the publication by Biswas [43]. The article titled "Tomato prices are on fire — and will not come down soon. Here is why" by P. Biswas in The Indian Express highlights the surge in tomato prices and the reasons behind it. According to the article published on June 29, 2023, the cost of tomatoes has witnessed a significant increase and is expected to remain high for an extended period.

To assess the adaptability (flexibility in adjusting to new data and maintaining performance) of the proposed method, we examined the results for tomatoes using real-time data from 2023, sourced from the Annual Report of the "Agricultural Produce Market Committee, Pune (Krushi Utpanna Bazar Samiti, Pune)" [44]. The maximum rate for tomatoes in July was 4500 INR, with an average rate of 3800 INR. In August, the maximum rate was 4000 INR, with an average rate of 3600 INR. For September, the maximum rate remained 4000 INR, with an average rate of 3500 INR. These results demonstrate that our proposed approach is adaptable.

These findings can empower stakeholders in the market, particularly farmers, to make informed decisions on optimal harvesting and crop cultivation strategies, maximizing their profits during these high-price periods. Overall, our research contributes valuable insights for better decision-making in the fruit and vegetable market in Maharashtra.

D. Future Scope

In terms of future prospects, our research aims to develop a farmer-centric application specifically tailored for predicting and assisting in crop planning within Maharashtra. This application will be continuously updated with the latest datasets, ensuring that farmers have access to the most relevant and accurate information for decision-making. Fig. 6 shows how application should be developed and work.

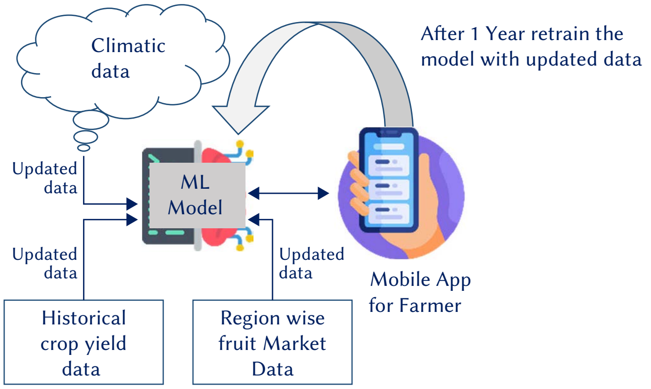

Climatic data
After 1 Year retrain the model with updated data
Updated data
ML Model
Updated data
Updated data
Mobile App for Farmer
Historical crop yield data
Region wise fruit Market Data

Fig. 6. Future working of Mobile Application for farmer.

By extending the scope of future application to encompass not just Maharashtra, but the entire nation of India and eventually the worldwide farming community, we have the thrilling chance to transform agricultural methods on a global level. Our future goal is to use cutting-edge technology and data-driven analysis to give farmers from different countries the tools and information necessary to enhance their agricultural operations and increase crop yields.

Moreover, a potential direction for future study and development is to improve the prediction capabilities of our algorithm by integrating supplementary features into the Regressor model. To enhance the accuracy and usefulness of the insights supplied to farmers, we may incorporate parameters such as weather conditions, soil nutrient levels, and geographical variances.

However, it is important to confront certain challenges, such as apprehensions over privacy and confidentiality, particularly when handling sensitive agricultural data. Considering that the data may be spread out across several sources and places, it becomes crucial to apply privacy-preserving mechanisms. An effective strategy to address these issues is by implementing federated learning, which involves training the model collectively using decentralized data sources while ensuring data privacy is maintained. Studying and analyzing the viability of such methods will be a crucial component of future research.

The future scope of our research encompasses the development of a comprehensive, farmer-centric application for predictive analytics and crop planning, with a focus on scalability, accessibility, and privacy preservation. Through the utilization of cutting-edge technology and inventive approaches, our goal is to make substantial contributions to the worldwide agricultural community and enable farmers to succeed in a progressively intricate and ever-changing environment.

1. Limitation of Research – Future Work

In future study, it is necessary to solve various limitations.

- Quality and availability of data: Gathering large and high-quality agricultural data, particularly in rural regions, poses a significant challenge.
- Privacy and Confidentiality: Safeguarding sensitive farmer data while ensuring its utility for analysis presents legal and ethical hurdles.

- Establishing a balance between complex prediction models and easily understandable interpretations is essential for building trust and comprehension among farmers.
- Scalability and generalization: Validating and adapting models to account for regional variances is necessary when extending them from local to global dimensions.
- Giving appropriate access to advanced technology such as cloud computing and mobile applications continues to pose financial and logistical difficulties.
- Encouraging widespread adoption among farmers and stakeholders requires overcoming challenges such as technology literacy and cultural disagreement.

To tackle these difficulties, it is necessary to have collaboration between different fields of study, use creative approaches, and maintain continuous involvement with farming communities.

V. Conclusion

The influence of technology is driving significant changes in every field worldwide, including the agricultural sector in India. To bolster its development and growth, the Indian farming industry requires more technological support. Accurate price prediction of agricultural products is crucial to ensure fair returns for farmers and to help them recover their investments. Our proposed method offers a valuable framework for predicting fruit and vegetable prices in the Maharashtra market, leveraging various ML and DL algorithms. This approach provides critical insights for decision-making in the Indian farming sector, empowering farmers and policymakers with data-driven cultivation strategies, distribution optimization, and effective marketing. The analysis and evaluation of several regression algorithms revealed XGBRegressor, Random Forest, and Decision Tree as the most suitable models, boasting high R2 scores close to 1 and low MSE and MAE. Future prospects include creating a farmer-centric application for forecasting and crop planning in Maharashtra, with regularly updated datasets. Scaling the application to cover India and the world represents an exciting opportunity to revolutionize global farming practices and benefit farmers across borders.

Data Availability

The data used for this study are available from the authors on request.

Conflicts of Interest

The authors have nothing to declare as conflicts of interest.

Funding Statement

The authors did not receive any financial support for conducting the research, writing the article, or publishing it.

References

[1] Q. Zhang, "Opinion paper: Precision agriculture, smart agriculture, or digital agriculture," *Computers and Electronics in Agriculture*, vol. 211, pp. 107982, 2023. doi:10.1016/j.compag.2023.107982.

[2] Y. Huang, Q. Zhang, "Agricultural Cybernetics", *Springer: Berlin/ Heidelberg, Germany*, 2021.

[3] Dvara Research, "Why don't Indian farmers grow more fruits and vegetables?," *Dvara Research Blog*. Jan. 30, 2013 [Online]. Available: https://www.dvara.com/research/blog/2013/01/30/why-dont-indian-farmers-grow-more-fruits-and-vegetables/

[4] J. Cheruku and V. Katekar, "Digitalisation of Agriculture in India: The case for doubling farmers' income," *Indian Institute of Public Administration*, pp. 194-205, 2023.

[5] M. Vibas and A. R. Raqueño, "A Mathematical Model for Estimating Retail Price Movements of Basic Fruit and Vegetable Commodities Using Time Series Analysis," *International Journal of Advance Study and Research Work*, vol. 2, no. 7, pp. 1–5, 2019. doi: 10.5281/zenodo.3333529.

[6] S. Rakhal and C. Brianne, "Price Transmission in Canadian Fresh Fruit Market: A Time Series Analysis", *International Journal of Food and Agricultural Economics (IJFAEC)*, vol. 9, no. 3, pp. 175-189, 2021. doi: 10.22004/ag.econ.313363.

[7] A. Jahangir, K. Jyoti, B. Deep Ji, and B. Anil, "Analysis of Prices and Arrivals of Apple Fruit in Narwal Market of Jammu", *Economic Affairs*, vol. 63, no. 1, pp. 107-111, March 2018, doi: 10.30954/0424-2513.2018.00150.13

[8] L. Nassar, I. E. Okwuchi, M. Saad, F. Karray and K. Ponnambalam, "Deep Learning Based Approach for Fresh Produce Market Price Prediction," *2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK*, 2020, pp. 1-7, doi: 10.1109/IJCNN48605.2020.9207537.

[9] I. Okwuchi, "Machine Learning based Models for Fresh Produce Yield and Price Forecasting for Strawberry Fruit," *M.S. thesis, Univ. of Waterloo*, 2020. [Online]. Available: http://hdl.handle.net/10012/15976.

[10] R. Agarwal and Prof. P. Sagar, "A Comparative Study of Supervised Machine Learning Algorithms for Fruit Prediction", *Journal of Web Development and Web Designing*, vol. 4, no. 1, pp. 14–18, Apr. 2019, doi: 10.5281/zenodo.2621205.

[11] R. Dharavath and E. Khosla, "Seasonal ARIMA to Forecast Fruits and Vegetable Agricultural Prices," *2019 IEEE International Symposium on Smart Electronic Systems (iSES) (Formerly iNiS), Rourkela, India*, 2019, pp. 47-52, doi: 10.1109/iSES47678.2019.00023.

[12] C. Sharma, R. Misra, M. Bhatia and P. Manani, "Price Prediction Model of fruits, Vegetables and Pulses according to Weather," *2023 13th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India*, 2023, pp. 347-351, doi: 10.1109/Confluence56041.2023.10048880.

[13] M. Kankar and M. A. Kumar, "Price Prediction of Agricultural Products Using Deep Learning," *Advanced Machine Intelligence and Signal Processing*, D. Gupta, K. Sambyo, M. Prasad, and S. Agarwal, Eds. *Singapore: Springer*, 2022, vol. 858, *Lecture Notes in Electrical Engineering*, pp. 495-506. doi: 10.1007/978-981-19-0840-8_38.

[14] R. K. Paul, M. Yeasin, P. Kumar, P. Kumar, M. Balasubramanian, H. S. Roy, et al., "Machine learning techniques for forecasting agricultural prices: A case of brinjal in Odisha, India," *PLoS ONE*, vol. 17, no. 7, p. e0270553, Jul. 2022, doi: 10.1371/journal.pone.0270553.

[15] C. Chai, J. Wang, Y. Luo, Z. Niu and G. Li, "Data Management for Machine Learning: A Survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 5, pp. 4646-4667, 1 May 2023, doi: 10.1109/TKDE.2022.3148237.

[16] Z. Luo, C. Fang, C. Liu and S. Liu, "Method for Cleaning Abnormal Data of Wind Turbine Power Curve Based on Density Clustering and Boundary Extraction," *IEEE Transactions on Sustainable Energy*, vol. 13, no. 2, pp. 1147-1159, April 2022, doi: 10.1109/TSTE.2021.3138757.

[17] F. Ridzuan and W. M. N. W. Zainon, "A Review on Data Cleansing Methods for Big Data," *Procedia Computer Science*, vol. 161, pp. 731-738, ISSN 1877-0509, 2019, doi: https://doi.org/10.1016/j.procs.2019.11.177.

[18] Y. Nieto, V. García-Díaz, C. Montenegro, C. C. González and R. González Crespo, "Usage of Machine Learning for Strategic Decision Making at Higher Educational Institutions," *IEEE Access*, vol. 7, pp. 75007-75017, 2019, doi: 10.1109/ACCESS.2019.2919343.

[19] D. P. Kumar, T. Amgoth, and C. S. R. Annavarapu, "Machine learning algorithms for wireless sensor networks: A survey," *Information Fusion*, vol. 49, pp. 1-25, 2019, doi: 10.1016/j.inffus.2018.09.013.

[20] Y. Nieto, V. García-Díaz, C. Montenegro, et al., "Supporting academic decision making at higher educational institutions using machine learning-based algorithms," *Soft Computing*, vol. 23, no. 12, pp. 4145-4153, 2019, doi: 10.1007/s00500-018-3064-6

[21] M. Ganesan, A. Suruliandi, S. P. Raja, and E. Poongothai, "An Empirical Evaluation of Machine Learning Techniques for Crop Prediction," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 4, pp. 96-104, 2023, doi: 10.9781/ijimai.2022.12.004.

[22] T. Ivanovski, G. Zhang, T. Jemrić, M. Gulić and M. Matetić, "Fruit

firmness prediction using multiple linear regression," *2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO), Opatija, Croatia*, 2020, pp. 1306-1311, doi: 10.23919/MIPRO48935.2020.9245213.

[23] D. Maulud and A. M. Abdulazeez, "A Review on Linear Regression Comprehensive in Machine Learning", *Journal of Applied Science and Technology Trends*, vol. 1, no. 2, pp. 140-147, Dec. 2020. doi: 10.38094/jastt1457

[24] A. Tsigler and P. L. Bartlett, "Benign overfitting in ridge regression," *Journal of Machine Learning Research*, vol. 24, no. 123, pp. 1-76, 2023. [Online]. Available: http://jmlr.org/papers/v24/22-1398.html

[25] H. Xu, C. Caramanis and S. Mannor, "Robust Regression and Lasso," *IEEE Transactions on Information Theory*, vol. 56, no. 7, pp. 3561-3574, July 2010, doi: 10.1109/TIT.2010.2048503.

[26] M. Kück and M. Freitag, "Forecasting of customer demands for production planning by local k-nearest neighbor models," *IEEE Transactions on Engineering Management*, vol. 231, p. 107837, 2021, ISSN: 0925-5273, doi: 10.1016/j.ijpe.2020.107837

[27] I. N. Yulita, A. S. Abdullah, A. Helen, S. Hadi, A. Sholahuddin, and J. Rejito, "Comparison multi-layer perceptron and linear regression for time series prediction of novel coronavirus covid-19 data in West Java," *Journal of Physics: Conference Series*, vol. 1722, no. 1, p. 012021, Jan. 2021. doi: 10.1088/1742-6596/1722/1/012021.

[28] H. Luo, F. Cheng, H. Yu and Y. Yi, "SDTR: Soft Decision Tree Regressor for Tabular Data," *IEEE Access*, vol. 9, pp. 55999-56011, 2021, doi: 10.1109/ACCESS.2021.3070575.

[29] E. Pekel, "Estimation of soil moisture using decision tree regression," *Theoretical and Applied Climatology*, vol. 139, no. 3, pp. 1111–1119, Mar. 2020, doi: 10.1007/s00704-019-03048-8.

[30] H. Wang, Q. Yilihamu, M. Yuan, H. Bai, H. Xu, and J. Wu, "Prediction models of soil heavy metal(loid)s concentration for agricultural land in Dongli: A comparison of regression and random forest," *Ecological Indicators*, vol. 119, p. 106801, 2020. doi: 10.1016/j.ecolind.2020.106801.

[31] M. Alida and M. Mustikasari, "Rupiah Exchange Prediction of US Dollar Using Linear, Polynomial, and Radial Basis Function Kernel in Support Vector Regression," *Jurnal Online Informatika*, vol. 5, no. 1, pp. 53-60, 2020. doi: 10.15575/join.v5i1.537

[32] C. R. Madhuri, G. Anuradha and M. V. Pujitha, "House Price Prediction Using Regression Techniques: A Comparative Study," *2019 International Conference on Smart Structures and Systems (ICSSS), Chennai, India*, 2019, pp. 1-5, doi: 10.1109/ICSSS.2019.8882834.

[33] Zhagparov, Z. Buribayev, S. Joldasbayev, A. Yerkosova and M. Zhassuzak, "Building a System for Predicting the Yield of Grain Crops Based On Machine Learning Using the XGBRegressor Algorithm," *2021 IEEE International Conference on Smart Information Systems and Technologies (SIST), Nur-Sultan, Kazakhstan*, 2021, pp. 1-5, doi: 10.1109/SIST50301.2021.9465938.

[34] A. Thaniserikaran, B. Sriphani Vardhan, A. Rahman Mateen Syed, M. Abdul Muqeet, A. Khot and B. K. Tejas, "The prediction of cern electron mass collision by using CATBoosting and LGBMR," *2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kharagpur, India*, 2022, pp. 1-5, doi: 10.1109/ICCCNT54827.2022.9984588.

[35] A. Botchkarev, "A new typology design of performance metrics to measure errors in machine learning regression algorithms," *Interdisciplinary Journal of Information, Knowledge, and Management*, vol. 14, pp. 045-076, 2019. doi: 10.28945/4184

[36] Chicco D, Warrens MJ, Jurman G. "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation." *PeerJ Computer Science*, vol. 7, p. e623, 2021.doi: 10.7717/peerj-cs.623

[37] Q. Sun, W. Zhou, and J. Fan, "Adaptive Huber Regression," *in Journal of the American Statistical Association*, vol. 115, no. 529, pp. 254-265, 2020. doi: 10.1080/01621459.2018.1543124

[38] M. Eppert, P. Fent, and T. Neumann, "A Tailored Regression for Learned Indexes: Logarithmic Error Regression," *in Fourth Workshop in Exploiting AI Techniques for Data Management (aiDM '21), Virtual Event, China*, 2021, pp. 9-15, doi: 10.1145/3464509.3464891

[39] L. F. Tratar and E. Strmčnik, "The comparison of Holt–Winters method and Multiple regression method: A case study," *Energy*, vol. 109, pp. 266-276, 2016. [Online]. Available: https://doi.org/10.1016/j.energy.2016.04.115

[40] S. Mirzaei, G.M. Borzadaran, M. Amini, and H. Jabbari, "A comparative study of the Gini coefficient estimators based on the regression approach," *Communications for Statistical Applications and Methods*, vol. 24, no. 4. *The Korean Statistical Society*, pp. 339–351, 31-Jul-2017. doi:10.5351/csam.2017.24.4.339.

[41] P. Rodríguez, M. A. Bautista, J. Gonzalez, and S. Escalera, "Beyond one-hot encoding: Lower dimensional target embedding," *Image and Vision Computing*, vol. 75, pp. 21-31, Apr. 2018. Doi: 10.1016/j.imavis.2018.04.00

[42] E. Bisong and E. Bisong,"Introduction to Scikit-learn," *in Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*, 2019, pp. 215-229. Doi: 10.1007/978-1-4842-4470-8_18

[43] P. Biswas, "Tomato prices are on fire — and will not come down soon. Here is why," *The Indian Express*, Online, June 29, 2023. [Accessed: July 20, 2023]. Available: https://indianexpress.com/article/explained/explained-economics/why-tomato-prices-high-8689168/

[44] Agricultural Produce Market Committee, Pune, "Annual Report of 2023 Agricultural Produce Market Committee, Pune", [Online] http://www.puneapmc.org/rates.aspx [Last Accessed: 30/04/2024].

Dr. Nilesh P. Sable

Dr. Nilesh P. Sable is a senior member IEEE and working as Associate Professor, Head Department of Computer Science & Engineering (Artificial Intelligence) at Vishwakarma Institute of Information Technology, Pune, India. He has completed his Ph.D. in Computer Science & Engineering from Kalinga University, Raipur. He has 16 + years of teaching and research experience. He is guiding 4 Ph.D. students in the area of Machine Learning, Federated Learning and IoT under his supervision from SPPU. He is working as Research Advisory Committee (RAC) Member for various Research Centres. He is a reviewer for various journals and conferences of the repute. He has published 75+ papers in National, International conferences and Journals. He had Filed and Published 16 Patents and 18 Copyrights. He has authored books published by National/International publishers. He is also the recipient of the "Distinguished Performance Award" by Vishwakarma Institute of Information Technology. He has delivered 50+ lectures at national and international level.

Rajkumar V. Patil

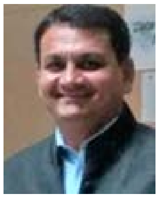
Rajkumar received his Bachelor of Engineering (Computer Engineering) in 2018 from Savitribai Phule Pune University, Pune, India, and his Master of Engineering in Computer Engineering in 2020 from Smt. Kashibai Navale College of Engineering, Pune, India. Cyber Physical systems for healthcare, Trust management, Machine learning, Web technologies, Algorithms and Explainable AI are his areas of interest. He has around 3 years of teaching and research experience. He has published research articles and book chapters in international journals. He is reviewer for several international journals. He is presently working as assistant professor at MIT Art, Design and Technology University, Pune and research scholar at Vishwakarma Institute of Information Technology Pune.

M. Deore

M. Deore is working as an Asst. Professor in Computer Engineering Department at MKSSS's Cummins College of Engineering for Women, Pune 411051, India. He was awarded his Master of Technology Degree from Bharati Vidyapeeth Deemed University College of Engineering, Dhankawadi, Pune. He received doctoral degree from Swami Ramanand Teertha Marathwada University, Nanded, India in 2020. He has 16 + years of teaching and research experience. He is a reviewer for various journals and conferences of the repute. He has published 25+ papers in National, International conferences and Journals. His areas of interest are big data, Security, Computer Networks and Machine learning. He has Fourteen years' experience in teaching.

Ratnmala Bhimanpallewar

Ratnmala Bhimanpallewar holds a Doctor of Philosophy degree in Computer Science and Engineering from K L University, Vijaywada, India. She has received her master's (M.E. Computer Science and Engineering) degree from PICT, Savitribai Phule Pune University, Pune, India. She is working as an Assistant Professor in the Information Technology Department of Vishwakarma Institute of Information Technology, Kondhwa (Bk.), Pune. She has 14 years of working experience. Her area of interest is Databases, Machine Learning and IoT She is a lifetime member of ISTE. She has completed the funded research project under SPPU ASPIRE scheme.



Parikshit N. Mahalle

Dr Parikshit is a senior member IEEE and is Professor, Dean Research and Development and Head - Department of Artificial Intelligence and Data Science at Vishwakarma Institute of Information Technology, Pune, India. He completed his Ph. D from Aalborg University, Denmark and continued as Post Doc Researcher at CMI, Copenhagen, Denmark. He has 23 + years of teaching and research experience. He is an ex- Board of Studies, Ex-Chairman at various Universities and autonomous colleges across India. He has 15 patents, 200+ research publications and authored/edited 60+ books with Springer, CRC Press, Cambridge University Press, etc. He is editor in chief for IGI Global –International Journal of Rough Sets and Data Analysis, Inter-science International Journal of Grid and Utility Computing, member-Editorial Review Board for IGI Global – International Journal of Ambient Computing and Intelligence and reviewer for various journals and conferences of the repute. His research interests are Machine Learning, Data Science, Cognitive Computing, Algorithms, Internet of Things, Identity Management and Security. He is guiding 8 PhD students in the area of IoT and machine learning and 6 students have successfully defended their PhD under his supervision from SPPU. He is also the recipient of the "Best Faculty Award" by Sinhgad Institutes and Cognizant Technologies Solutions and State Level Meritorius Teacher Award. He has delivered 200 + lectures at national and international level.

Regular Issue

Anti-Diabetic Therapeutic Medicinal Plant Identification Using Deep Fused Discriminant Subspace Ensemble (D^{2}SE)

N. Sasikaladevi[1], S. Pradeepa[2], A. Revathi[3], S. Vimal[4]*, Gaurav Dhiman[5,6,7]*

[1] Department of Computer Science and Engineering, SASTRA Deemed University, Thanjavur-613 401, Tamil Nadu (India)
[2] Department of Information Technology, SASTRA Deemed University, Thanjavur -613 401, Tamil Nadu (India)
[3] Department of Electrical and Computer Engineering, SASTRA Deemed University, Thanjavur -613 401, Tamil Nadu (India)
[4] Department of Artificial Intelligence & Data Science, Sri Eshwar College of Engineering, Coimbatore, Tamil Nadu (India)
[5] Department of Electrical and Computer Engineering, Lebanese American University, Byblos (Lebanon)
[6] Centre of Research Impact and Outcome, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab (India)
[7] MEU Research Unit, Middle East University, Amman (Jordan)

* Corresponding author: svimalphd@gmail.com (S.Vimal), gdhiman0001@gmail.com (G. Dhiman)

Received 29 March 2023 | Accepted 1 September 2023 | Published 27 May 2024

LA UNIVERSIDAD
EN INTERNET

Abstract

About 422 million people worldwide have diabetes, the majority living in low-and middle-income countries, and 1.5 million deaths are directly attributed to diabetes each year. According to the Botanical Survey of India, India is home to more than 8,000 species of medicinal plants. The natural medications with antidiabetic activity are widely formulated because they have better compatibility with human body, are easily available and have less side effects. They may act as an alternative source of antidiabetic agents. The fused deep neural network (DNN) model with Discriminant Subspace Ensemble is designed to identify the diabetic plants from VNPlant200 data set. Here, the deep features are extracted using DenseNet201 and the matrix-based discriminant analysis is adopted to learn the discriminative feature subspace for classification. To further improve the performance of discriminative subspace, a nearest neighbors technique is used to produce a subspace ensemble for final diabetic therapeutic medicinal plant image classification. The developed model attained the highest accuracy of 97.5% which is better compared to other state of art algorithms. Finally, the model is integrated into a mobile app for robust classification of anti-diabetic therapeutic medicinal plant with real field images.

Keywords

Classification, Deep Learning, Diabetic Plant Identification, Discriminant Subspace Ensemble, Internet Of Things (IoT).

DOI: 10.9781/ijimai.2024.05.003

I. Introduction

In 2021, approximately 537 million adults (20-79 years) are living with diabetes. The total number of people living with diabetes is projected to rise to 643 million by 2030 and 783 million by 2045.3 in 4 adults with diabetes live in low- and middle-income countries. Almost 1 in 2 (240 million) adults living with diabetes are undiagnosed. Diabetes caused 6.7 million deaths. Diabetes caused at least USD 966 billion dollars in health expenditure – 9% of total spending on adults. More than 1.2 million children and adolescents (0-19 years) are living with type 1 diabetes. 1 in 6 live births (21 million) are affected by diabetes during pregnancy. 541 million adults are at increased risk of developing type 2 diabetes. Antidiabetic herbal formulations are considered to be more effective for treatment of diabetes. A high number of plants and plant products have been scientifically tested and reported to possess anti diabetic activity [1].

Antidiabetic herbal formulations are considered to be more effective for treatment of diabetes. The global worsening of morbidity and mortality from diabetes [2] [3] justifies the need for more diversified research for new therapies. Throughout human history, medicinal plants have been used for the prevention and treatment of both human and animal diseases [4] [5]. Medicinal plants have been recognized as a stable source for drug discovery since ancient times [6] [7] and The World Health Organization has reported an increased patronage of natural and medicinal plant drug products[1].

[1] https://www.who.int/health-topics/diabetes

Please cite this article as:
N. Sasikaladevi, A. Revathi, S. Pradeepa, S.Vimal, G. Dhiman. Anti-Diabetic Therapeutic Medicinal Plant Identification Using Deep Fused Discriminant Subspace Ensemble (D^{2}SE), International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 1, pp. 55-65, 2024, http://dx.doi.org/10.9781/ijimai.2024.05.003

Many modern drugs are obtained from medicinal plants and further purified or optimized using structure-activity relationship-driven drug design and pharmacokinetic parameters [8] [9]. Evidence-based application of phytochemicals from plants in the management of diseases has received wide acceptability [10]. For example, several reports of medicinal plants with anticancer activities have been published [11] [12]. Ethnopharmacological surveys of plants and phytochemicals with antihypertensive activities [13] have been well documented.

There is also substantial literature of their utility in treatment of other chronic diseases such as Alzheimer's [14], depressive disorders [15], Parkinson's disease [16] and diabetes [17]. Various plants and plant parts have been investigated for their hypoglycemic activities as potential medicine in the treatment of diabetes mellitus [18]. By way of examples, phytocompounds from the fruit of Momordica charantia (bitter lemon) have been extensively studied for antidiabetic effects [19].

The roots of Zingiber officinale (ginger) exert antidiabetic and hypolipidemic effects on streptozotocin-induced diabetic rats [20]. Bidens pilosa has been shown to reduce fasting blood glucose level and hemoglobin A1c (HbA1c) in clinical trials [15]; three variants of B. pilosa were shown to possess anti-diabetic properties [21]. The hydroethanolic extract of the seed of Parinaricuratellifolia reduces plasma glucose levels and low-density lipoproteins in diabetic rats [22].

The blood sugar reducing effects of Gymnemasylvestre popularly known as 'gurmar' ('sugar destroyer') has been widely studied. Phytochemical constituents of Glycyrrizauralensis (licorice) have been found to exhibit profound antidiabetic properties in experimental animals [23]. While some studies do consider the potential molecular or cellular mechanisms of the antidiabetic effects [24], others focus on potential properties such as antioxidant [5] and anti-obesity [25] effects without direct discussion of mechanism.

Modern medicine is massively produced for medical treatment, but many countries are now opting for traditional medicine due to the limitation of synthetic drugs in controlling and curing chronic diseases (WHO, 1999). Traditional medicines are used extensively in the pharmaceutical industry, where a quarter of the globally prescribed drug is extracted from medicinal plants. This is due to the benefits of medicinal plants that offer substantially lower adverse reactions and are more cost-effective than synthetic drugs. Furthermore, bioactive compounds such as phenolic, carotenoids, anthocyanin, and tocopherols that can be extracted from medicinal plants [26] serve as antioxidants, anti-allergenic, anti-inflammatory, antibacterial, and also anti-hepatotoxic.

Nonetheless, the task of manually identifying medicinal plants is complicated and time-consuming, similar to other plant recognition, due to the unavailability of expert opinions [23] [24] [25]. Inspired by these problems, researchers introduced numerous automatic plant or leaf recognition systems, most of which utilized Machine learning and deep learning approaches.

Deep learning methods play an essential role in plant leaf classifications. Deep learning methods are based on Convolutional Neural Networks (CNN), which comprise deep feature extractors. When the system is developed to classify medicinal plants, that system should be trained by a large dataset, including noisy images. The primary goal of the recognition system is to interpret the leaf images with high accuracy compared to the manual task. In a highly densely populated country like India, there is a lack of manual experts. Hence, there is a demand to accurately develop an AI-based tool for medicinal plants.

This paper is organized as follows. Section II reviews the existing machine learning algorithms and deep learning methods for medicinal plant classification that are categorized according to their performance. Section III describes the proposed D2SE model for medicinal plant leaf classification and elucidates the details of the medicinal plants which are used in this work. Section IV elaborates on the proposed system's experimental work and performance, and Section V concludes the paper.

II. Related Works

CNN-based medicinal plant identification is proposed [27] and attained an accuracy of 88.26%. Amuthalingeswaran et al. [28] proposed Deep Neural Network (DNN) model for the medicinal leaf classification for four medicinal plants such as Catharanthus Roseus, Tephrosia Purpurea, Phyllanthus amarus, and Abutilon Indicum. Putri et al. [29] proposed the CNN model for Indonesian medicinal plant leaf identification. Dileep et al. [30] applied the Alexnet neural network model to medicinal plant leaf for deep feature extraction, and then the Support Vector Machine (SVM) classifier is applied for classification. ResNet-based medicinal plant classification is proposed [31] Liu et al. [32] applied GoogleNet for the classification of Chinese herbal plant classification. Muneer et al. [33] applied the SVM and DNN classifier to Malaysian herbal leaf data. Mukherjee et al. [34] proposed a CNN model with binary particle swarm optimization based hyperparameter tuning method for medicinal plant classification. The Multiorgan-based classification model is proposed by Lee et al. [35]. All the diagnosis methods are tabulated in Table I.

Deep learning models for the classification of diabetic medicinal plants are limited. Furthermore, Table I shows that most of the current models yield low accuracy, which is insufficient for robust identification. Hence, there is a demand for developing a robust deep learning model for the classification of diabetic therapeutic medicinal plant.

III. Proposed D^{2}SE Framework

The deep learning phase of D^{2}SEhas three primary stages: data augmentation, feature extraction using Densenet201 and Matrix-Based Discriminative Feature Subspace Learning. Fig. 1 shows proposed D^{2}SE framework for Anti-diabetic therapeutic medicinal plant identification.

A. Data Collection and Preprocessing

Applications of image processing and computer vision techniques for identifying medicinal plants are critical, as many of them are under extinction, per the International Union for Conservation of Nature (IUCN) records. Hence, the digitization of valuable medicinal plants is crucial for the conservation of biodiversity. Studies reveal that building an intelligent system for recognizing medicinal herbs requires a decent size of the plant image dataset. Thus, diabetic related images of VNPlant-200 dataset are acquired from the National Institute of Medicinal Materials. They are labeled manually by botanists and technical expert. This dataset is more challenging and noisier than others because many leaves appear together in a single image and it also contains the background such as soil, treebark, flower, etc. Table II illustrates several image examples from diabetic class ofVNPlant-200 dataset, with two versions of resolutions: 256X 256 and 512X512 pixels.

This dataset is divided into different partitions for training and Validation. In order to perform hold-out Validation, a set of images is to be kept aside for training. Those unseen data are to be given to the model for perfect Validation. Hence, the dataset is divided into 70%for training and 30% for Validation. Table III shows the sample references for the medicinal plant available in the VNPlant 200 data set.

B. Feature Extraction Using Densenet201

Hyper-tuning of Deep transfer learning models (DTL) can improve results in classification problems [46]-[49]. Here, a Deep Transfer

TABLE I. State-of-the-art Methods for the Medicinal Plant Classification

Authors	Method used	Dataset	Performance	Drawback
(Quoc, 2020) [27]	VGG16, Resnet50, InceptionV3, DenseNet121, Xception and MobileNet	Vietnamese medicinal plant images	Accuracy of 88.26%	Low accuracy
(Amuthalingeswaran, 2019) [28]	Deep neural network	Private dataset with four plants	Accuracy 85%	Low accuracy and four-class classifier only
(Putri, 2021) [29]	CNN	Indonesian medicinal leaf dataset	Not given	Nonstandard analysis
(Dileep, 2019) [30]	Alexnet+SVM	Kerala medicinal plant leaf dataset	Accuracy 96.76%	Slow convergence
(Liu, 2021) [31]	Resnet	Grassland plant data set	Accuracy 96.8%	Nonstandard data division
(Liu, 2018) [32]	GoogleNet	Chinese herbal medicine	Accuracy 62.8%	Low accuracy
(Muneer,2020) [33]	SVM and DNN	Medical herbs in Malaysia	Accuracy with SVM: 74.63% Accuracy with DNN:93%	Low accuracy
(Mukherjee, 2021) [34]	CNN with hyperparameter tuning	Private dataset with Neem, Tulsi, and Kalmegh	Accuracy 99%	Nonstandard database. Limited class classification
(Tiwari, 2022) [36]	DNN	Private plant leaf dataset	Accuracy 97.67%	Nonstandard analysis
(Lee, 2018) [35]	CNN	Multiorgan dataset	Leaf: 76% Flower: 81%	Low precision
(Bhuiyan, 2021) [37]	CNN	Bangladesh medicinal plant	Accuracy of 84%	Low Accuracy
(Almazaydeh, 2022) [38]	R-CNN	Mendeley medicinal plant dataset	Accuracy of 95.7%	slow convergence

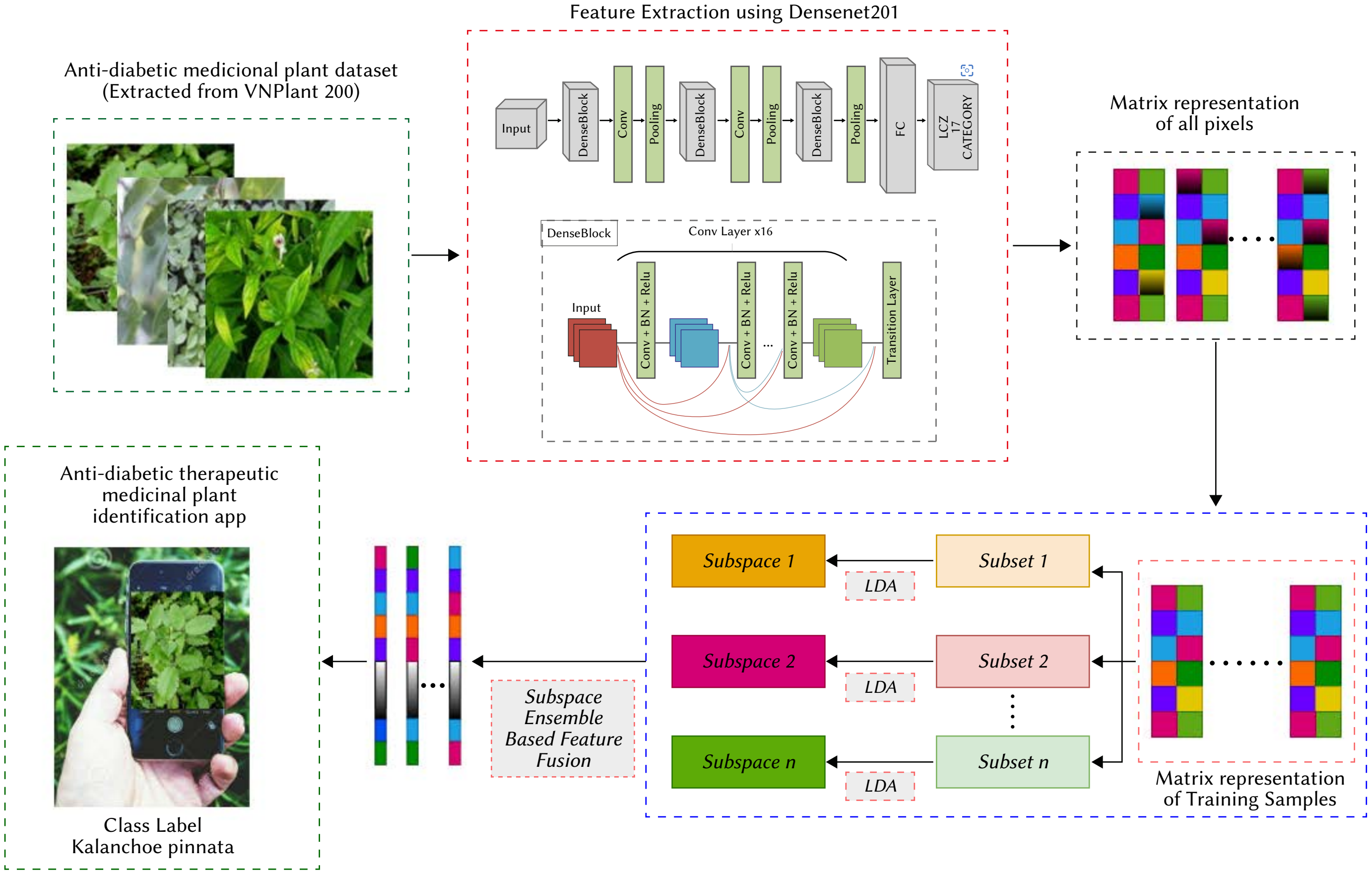
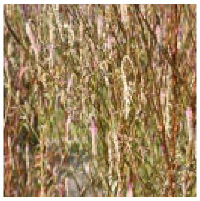
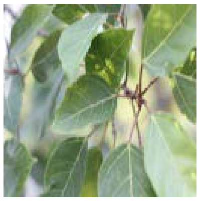
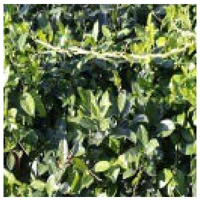
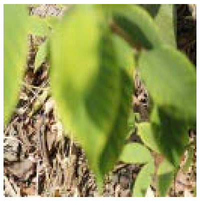
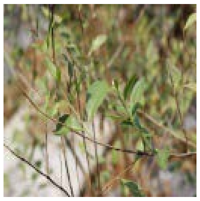

Feature Extraction using Densenet201
Anti-diabetic medicinal plant dataset (Extracted from VNPlant 200)
Input
DenseBlock
Conv
Pooling
DenseBlock
Conv
Pooling
DenseBlock
Pooling
FC
LCZ 17 CATEGORY
DenseBlock
Conv Layer x16
Input
Conv + BN + Relu
Conv + BN + Relu
...
Conv + BN + Relu
Transition Layer
Matrix representation of all pixels
Anti-diabetic therapeutic medicinal plant identification app
Class Label
Kalanchoe pinnata
Subspace Ensemble Based Feature Fusion
Subspace 1
Subspace 2
Subspace n
LDA
LDA
LDA
Subset 1
Subset 2
Subset n
Matrix representation of Training Samples

Fig. 1. Proposed D²SE framework for Anti-diabetic therapeutic medicinal plant identification.

TABLE II. D^{2}SE – Medicinal Plants and Their Medicinal Values

 Abrusprecatorius	 Aloe vera	 Andrographis paniculata	 Caesalpinia sappan	 Capsicum annuum	 Carica papaya
 Catharanthus roseus	 Celosia argentea	 Centella asiatica	 Coixlacryma-jobi	 Costusspeciosus	 Curculigoorchioides
 Euphorbia hirta	 Ficus auriculata	 Ficus racemosa	 Glycosmis pentaphylla	 Gymnemasylvestre	 Holarrhenapubescens
 Kalanchoe pinnata	 Lawsoniainermis	 Mangifera	 Melastomamalabathricum	 Mentha Spicata	 Mimosa pudica
 Morindacitrifolia	 Moringa oleifera	 Morus alba	 Nelumbo nucifera	 Ocimumbasilicum	 Ocimumgratissimum
 Ocimum sanctum	 Psidium guajava	 Senna alata	 Tabernaemontanadivaricata	 Tamarindus indica	 Zingiber officinale

Learning model with DenseNet201 is used for feature extraction by using its own learned weights on the Anti-diabetic medicinal plant dataset (Extracted from VNPlant 200). The DenseNet201 exploits the condensed network providing easy to train and highly parametrically efficient models due to the possibility of feature reuse by different layers which increases variation in the subsequent layer input and improves the performance. Fig. 2 shows the feature extraction techniques using DenseNet201.

The feature concatenation can be mathematically explained as in Equation (1).

$$Z^l = H_l([Z^0, Z^1, \ldots\ldots, Z^{l-1}]) \tag{1}$$

Here, H_l is a non-linear transformation which can be defined as a composite function comprising of batch normalization (BN), followed by a rectified linear unit function (ReLU) and a Convolution of (3X3). $[Z^0, Z^1, ..., Z^{l-1}]$ refers to the feature map concatenation corresponding to layer 0 to 1−1 which are integrated in a single tensor for ease of implementation. For down-sampling purposes, dense blocks are created in the network architecture which are separated by layers called transition layers which consist of BN followed by 1 × 1 convolution layer and finally a 2 × 2 average pooling layer. DenseNet201 performs sufficiently well even with a smaller growth rate owing to its architecture where feature maps are considered as a global state of the network. Therefore, each successive layer has access to all feature maps of preceding layers. k feature maps are added to

TABLE III. Sample References for the Medicinal Plant Available in the Vnplant-200 Data Set

Indian Medicinal Plant	Plant Part	Therapeutic use identifier	Reference
Abrusprecatorius	leaf	MESH:D003920, UMLS:C0011849, DOID:9351, ICD-11:5A14	(Kirtikar. 2012) [39]
Aloe vera	leaf	MESH:D003920, UMLS:C0011849, DOID:9351, ICD-11:5A14	(NISC, 2004) [40]
Andrographis paniculata	whole plant	MESH:D003920, UMLS:C0011849, DOID:9351, ICD-11:5A14	(Kirtikar. 2012) [39]
Caesalpinia sappan	wood	MESH:D003920, UMLS:C0011849, DOID:9351, ICD-11:5A14	(Kirtikar. 2012) [39]
Capsicum annuum		MESH:D003920, UMLS:C0011849, DOID:9351, ICD-11:5A14	(Kirtikar. 2012) [39]
Carica papaya	Fruit, root, seed	MESH:D003920, UMLS:C0011849, DOID:9351, ICD-11:5A14	(NISC, 2004) [40]
Catharanthus roseus	Flower, leaf, plant exudate, root, whole plant	MESH:D003920, UMLS:C0011849, DOID:9351, ICD-11:5A14	(NISC, 2004) [40]
Celosia argentea	seed	MESH:D003920, UMLS:C0011849, DOID:9351, ICD-11:5A14	(Kirtikar. 2012) [39]
Centella asiatica		MESH:D003920, UMLS:C0011849, DOID:9351, ICD-11:5A14	(Sivarajan, 2017) [41]
Coixlacryma-jobi	Leaf, seed	MESH:D003920, UMLS:C0011849, DOID:9351, ICD-11:5A14	(Asolkar, 2010) [42]
Costusspeciosus		MESH:D003920, UMLS:C0011849, DOID:9351, ICD-11:5A14	(Sivarajan, 2017) [41]
Curculigoorchioides	Rhizome, root	MESH:D003920, UMLS:C0011849, DOID:9351, ICD-11:5A14	(Kirtikar. 2012) [39]
Euphorbia hirta	aerial part, flower, leaf	MESH:D003920, UMLS:C0011849, DOID:9351, ICD-11:5A14	(Kirtikar. 2012) [39]
Ficus auriculata	Fruit, leaf	MESH:D003920, UMLS:C0011849, DOID:9351, ICD-11:5A14	(Kirtikar. 2012) [39]
Ficus racemosa	Bark, fruit, leaf, plant exudate, root, seed, stem, whole plant	MESH:D003920, UMLS:C0011849, DOID:9351, ICD-11:5A14	(Kirtikar. 2012) [39] (NISC, 2004) (CSIR, 2010) [40] (Chopra, 2022) [43]
Glycosmis pentaphylla	leaf	MESH:D003920, UMLS:C0011849, DOID:9351, ICD-11:5A14	(Kirtikar. 2012) [39]
Gymnemasylvestre	Bark, fruit, leaf, root, shoot, stem, whole plant	MESH:D003920, UMLS:C0011849, DOID:9351, ICD-11:5A14	(Khare, 2008) (Chopra, 2022) [44] [43] (NISC, 2004) (Asolkar, 2010) [40] (CSIR, 2010) [45]
Holarrhenapubescens	Bark, leaf, root, seed,	MESH:D003920, UMLS:C0011849, DOID:9351, ICD-11:5A14	(Kirtikar. 2012) [39]
Kalanchoe pinnata	leaf	MESH:D003920, UMLS:C0011849, DOID:9351, ICD-11:5A14	(Khare, 2008) [44] (Asolkar, 2010) [42]
Lawsoniainermis	flower	MESH:D003920, UMLS:C0011849, DOID:9351, ICD-11:5A14	(Kirtikar. 2012) [39]
Mangifera	Bark, flower, fruit, leaf, seed	MESH:D003920, UMLS:C0011849, DOID:9351, ICD-11:5A14	(Khare, 2008) [44], (Kirtikar. 2012) [39] (NISC, 2004) [40]
Melastomamalabathricum	leaf	MESH:D003920, UMLS:C0011849, DOID:9351, ICD-11:5A14	(Kirtikar. 2012) [39]
Mentha Spicata	leaf	MESH:D003920, UMLS:C0011849, DOID:9351, ICD-11:5A14	(Kirtikar. 2012) [39]
Mimosa pudica	Leaf, root, whole plant	MESH:D003920, UMLS:C0011849, DOID:9351, ICD-11:5A14	(Kirtikar. 2012)[39]
Morindacitrifolia	fruit	MESH:D003920, UMLS:C0011849, DOID:9351, ICD-11:5A14	(Kirtikar. 2012) [39]
Moringa oleifera	aerial part, bark, fruit, leaf, seed	MESH:D003920, UMLS:C0011849, DOID:9351, ICD-11:5A14	(Kirtikar. 2012) [39]
Morus alba	leaf	MESH:D003920, UMLS:C0011849, DOID:9351, ICD-11:5A14	(Kirtikar. 2012) [39]
Nelumbo nucifera	Rhizome, seed, shoot	MESH:D003920, UMLS:C0011849, DOID:9351, ICD-11:5A14	(Kirtikar. 2012) [39]
Ocimumbasilicum		MESH:D003920, UMLS:C0011849, DOID:9351, ICD-11:5A14	(Kirtikar. 2012) [39]
Ocimumgratissimum	leaf	MESH:D003920, UMLS:C0011849, DOID:9351, ICD-11:5A14	(Kirtikar. 2012) [39]
Ocimum sanctum	aerial part, leaf, seed	MESH:D003920, UMLS:C0011849, DOID:9351, ICD-11:5A14	(Kirtikar. 2012) [39]
Psidium guajava	leaf	MESH:D003920, UMLS:C0011849, DOID:9351, ICD-11:5A14	(Khare, 2008) [44]
Senna alata	flower	MESH:D003920, UMLS:C0011849, DOID:9351, ICD-11:5A14	(Kirtikar. 2012) [39]
Tabernaemontanadivaricata	leaf	MESH:D003920, UMLS:C0011849, DOID:9351, ICD-11:5A14	(Kirtikar. 2012) [39]
Tamarindus indica	seed	MESH:D003920, UMLS:C0011849, DOID:9351, ICD-11:5A14	(Kirtikar. 2012) (Warrier, 2022) [46]
Zingiber officinale		MESH:D003920, UMLS:C0011849, DOID:9351, ICD-11:5A14	(Sivarajan, 2017) [41]

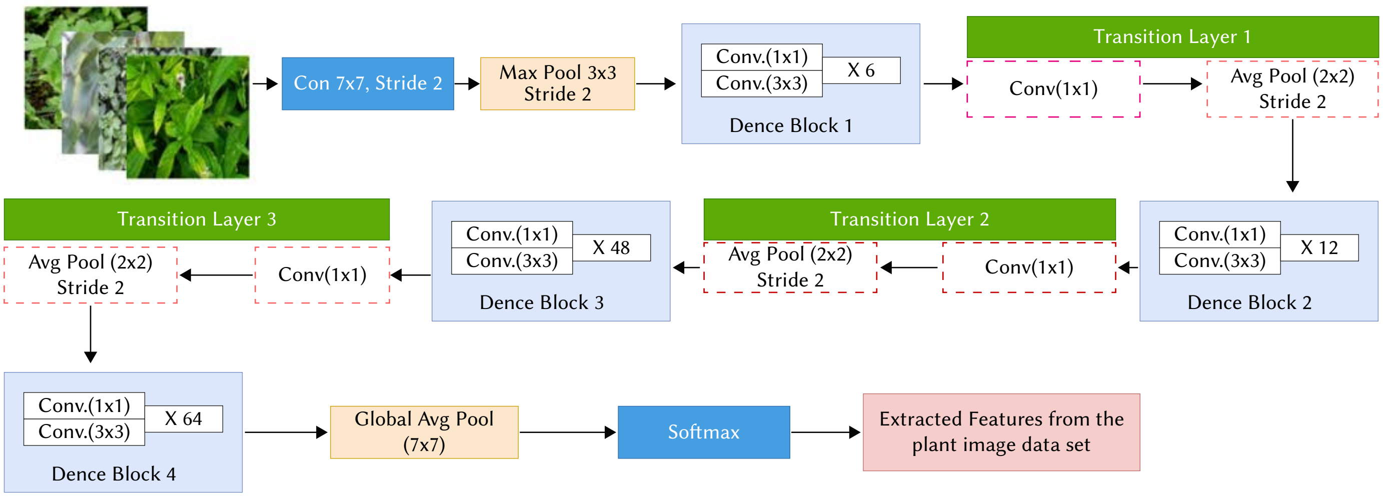


Fig. 2. Feature Extraction using DenseNet201.

the global state by each layer where the total number of input feature maps at lth layers (FM)lis defined in Equation (2).

$$(FM)^l = k^0 + k(l-1) \tag{2}$$

Here, the channels in the input layer is given by k^0. To improve computational efficiency, a1X1 convolution layer is introduced preceding every 3X3 convolution layer which decreases the number of input feature maps which are typically more than the output feature maps k. The feature extraction network followed by sigmoid activation is used in the classical DenseNet201 architecture.Each neuron in the fully connected dense layer is fully connected to each neuron in previous layer. It can be mathematically explained by a fully connected layer l whose input 2D feature map is expanded to a 1D feature vector as (3)-(5).

$$t^{l-1} = Bernoulli(p) \tag{3}$$

$$x^{l-1} = t^{l-1} * c^{l-1} \tag{4}$$

$$x^l = f(w^k x^{l-1} + o^l) \tag{5}$$

Here, Bernoulli function randomly generates a vector t^{l-1} obeying the 0−1 distribution with a specified probability. The vector dimension is c^{l-1}. The first two layers of the full connection layer use the dropout strategy to randomly block certain neurons according to a specified probability, which effectively prevents the over-fitting phenomena in the deep networks. w^l and o^l define the weighting and offset parameters of the fully connected layer respectively. The sigmoid activation function is to convert the non-normalized outputs to binary outputs as 0 or 1. The sigmoid function can be defined in Equation (6).

$$y = \frac{1}{1+e^{-(\sum w_i x_i)}} \tag{6}$$

where y is the output of the neuron. w_i and x_i represent the weights and inputs, respectively.

C. Matrix-Based Discriminative Feature Subspace Learning

The extracted features are represented by $X \in R^{m\times n}$. The Latent Dirichlet Allocation (LDA) is used to reduce the redundant information in this feature matrix and improve the discriminant ability of feature representation. LDA map X into a new subspace, in which the intraclass scatter is minimized and the interclass scatter is maximized.

Assume that $M = [\alpha_1, \alpha_2,.... \alpha_f] \in R^{m\times j}$ and $G = [v_1, v_2,.... v_c] \in R^{n\times c}$ are two transformation matrices. Then, the projection of X onto the $(f\times c)$ dimensional space $M \otimes G$ can be represented in Equation (7).

$$Z = M^T X G \tag{7}$$

Suppose there are P training pixels and L different classes to be classified. The j^{th} training pixel is X_j, where $j \in \{1, 2, ..., P\}$.

$F_W^G = \sum_{i=1}^{L} \sum_{x_k \in \pi_i} (X_k - \mu_i)\, GG^T (X_k - \mu)^T$ and μ_i denote the mean of all training pixels and training pixels in class Π_i, $i \in \{1, 2, ..., L\}$ respectively. P_i is the number of training pixels in class Π_i. To get the optimal projection matrices, we maximize the ratio of the interclass scatter matrix and the intraclass scatter matrix for LDA. Fortunately, the total scatter of the projected pixels can be measured by the trace of their covariance matrix. Therefore, the objective function can be described in Equation (8).

$$J = \frac{F_B}{F_W} \tag{8}$$

where F_B denotes the interclass scatter matrix, and F_W denotes the intraclass scatter matrix with the following definitions:

$$F_B = tr(\sum_{i=1}^{L} P_i\, (\mu_i - \mu)(\mu_i - \mu)^T) \tag{9}$$

$$F_w = tr(\sum_{i=1}^{L} \sum_{z_k \in \pi_i} (z_k - \mu_i)(z_k - \mu_i)^T) \tag{10}$$

$$\mu_i = \frac{1}{P_i} \sum Z_k = \frac{1}{P_i} \sum Z = M^T X_k G \tag{11}$$

The same as in [33], we utilize an iterative algorithm to optimize M and G. In particular, for a fixed G, (8) can be rewritten in the following form (Refer Equation (12)-(14)):

$$J = \frac{M^T F_B^G M}{M^T F_W^G M} \tag{12}$$

Where,

$$F_B^G = \sum_{i=1}^{L} P_i (\mu_i - \mu) GG^T (\mu_i - \mu)^T \tag{13}$$

$$F_W^G = \sum_{i=1}^{L} \sum_{x_k \in \pi_i} (X_k - \mu_i)\, GG^T (X_k - \mu)^T \tag{14}$$

Hence, the optimal solution of M consists of r eigenvectors corresponding to r maximal eigenvalues by computing an eigen decomposition on $(F_W^G)^{-1} F_B^G$. Subsequently, the optimal solution of G can be obtained when M is fixed. The whole process will iterate until a predefined convergence condition is arrived.

D. Learning Subspace Ensemble With Random Sampling

Ensemble learning has been proved to be an efficient technique to improve the stability and accuracy of single weak classifier by training several different classifiers and combining their decisions.

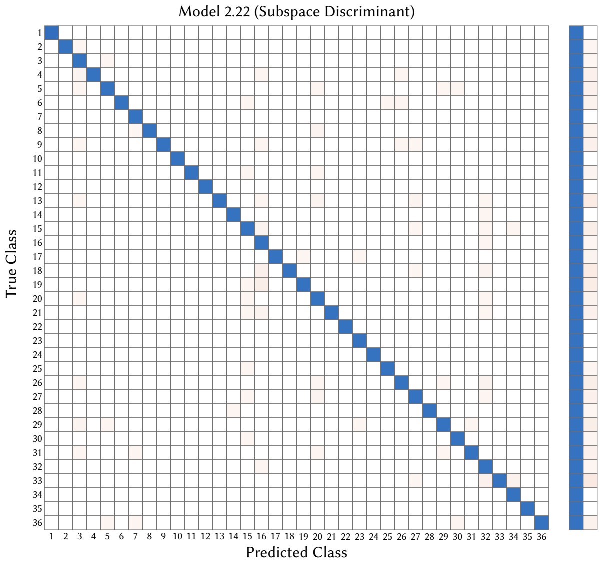


Fig. 3. D²SE- Confusion Matrix.

The core idea is to randomly sample a few training subsets from the original training set repeatedly with replacement. In this situation, the intraclass scatter matrix will be dominated by the majority classes. Thus, the dimensionality reduced subspace can preserve the structure information of majority classes while ignoring that of the minority classes, resulting in a decreased performance for subsequent classification. Here, we propose a random sampling- based subspace ensemble method for features extracted from the LDA. On one hand, motivated by bagging methods. random sampling repeatedly derives a few different training subsets, leading to diverse discriminant subspaces. These subspaces ensemble will improve the classification capacity. The bagging algorithm has been used to construct nearest neighbors ensemble for diabetic plant classification, and the experiments suggest that an optimized ensemble method could lead to improved results.

Fig. 1 shows the detailed flowchart of the proposed method. Suppose that we have a training set X, and X_j represents the j^{th} training pixel. First, n training subsets are randomly sampled. from X with replacement according to a certain proportion (e.g., 80%), and $X(i)$, $i \in \{1, 2, ..., n\}$ denotes the ith training subset. Then, for every subset $X(i)$, LDA is used to learn the optimal transformations $M(i)$ and $G(i)$ as discussed in Section III-C. Thereafter, for every couple $M(i)$ and $G(i)$, the projection matrix

$Z_j^{(i)}$ of X_j can be achieved according to (1). Therefore, we can obtain n projection matrices $\left\{Z_j^{(i)}, Z_j^{(2)} \ldots.. Z_j^{(n)}\right\}$. Finally, all of these matrices are reshaped to vectors, respectively. In feature fusion phase, we concatenate these vectors into one stacked vector, and Principal Component Analysis (PCA) is used to reduce the redundancy in the vector.

IV. Results

This section focuses on the experimental setup for validating the proposed model, followed by a performance analysis. Finally, the proposed D²SE is compared with other state-of-the-art deep learning models for classifying anti diabetic medicinal plants.

A. Experimental Setup and Evaluation

The proposed transfer learning model is implemented in MATLAB 2020b with GPU NVIDIA and 16GB RAM Workstation computer. VNPlant-200 is accepted as the benchmark data set by the researchers. The proposed model is trained and tested with the same dataset.

The Evaluation of the proposed model is based on two phases: training and Validation. The acceptable tuned hyperparameters are stablished to train the proposed transfer learning model, and the overall performance is validated. The first standard metric for evaluating the proposed model is based on confusion matrix, including sensitivity, specificity, precision, and accuracy. The confusion matrix is shown in Fig. 3 and the Parallel Co-ordinate Plot for Ensemble Subspace discriminant is depicted in Fig. 4.

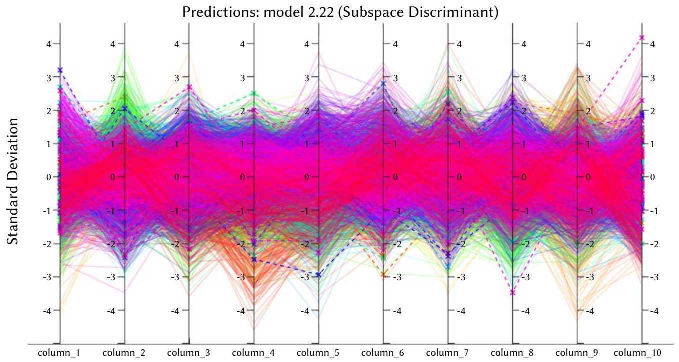


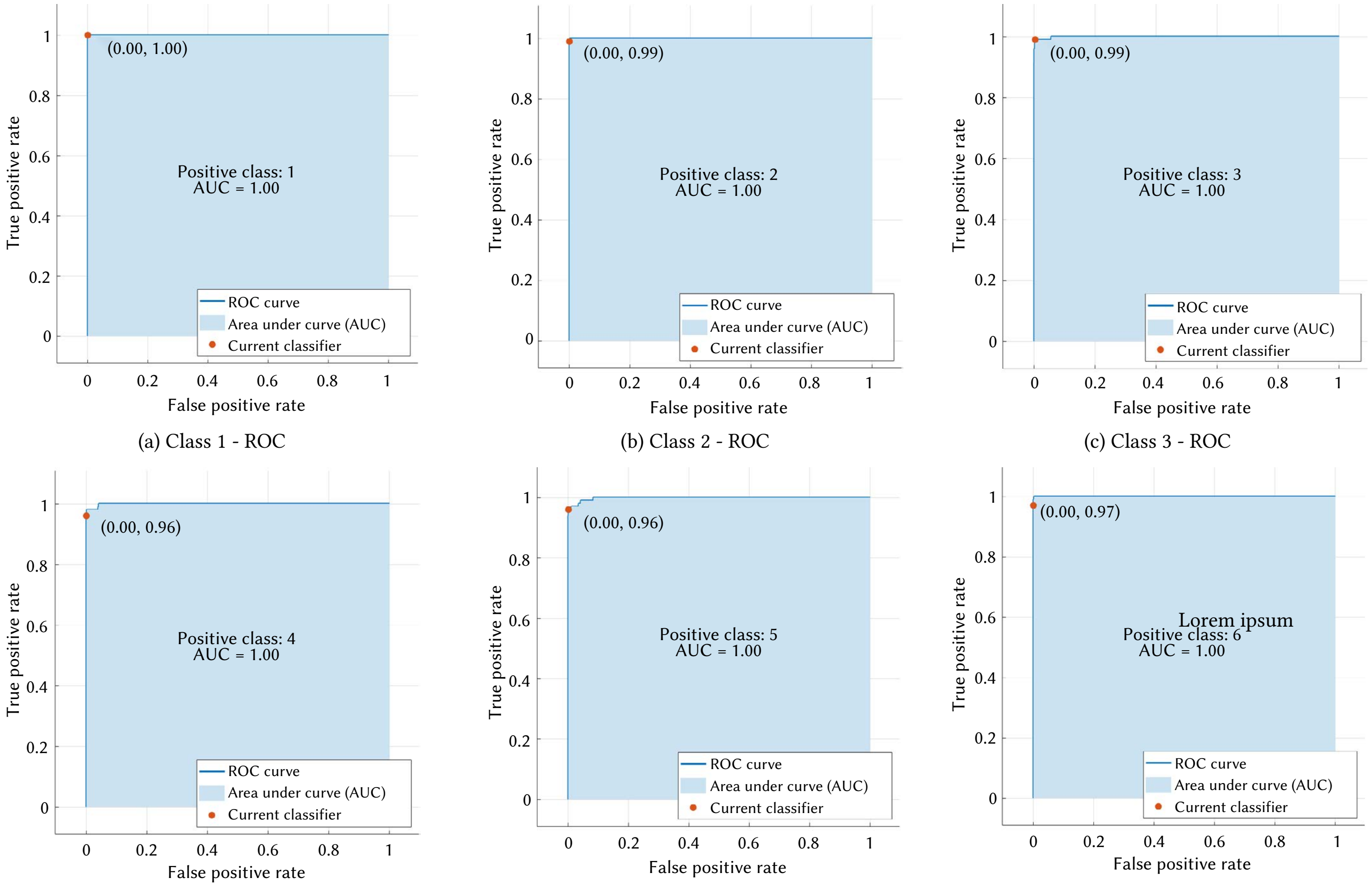
Fig. 4. Parallel Co-ordinate Plot – Ensemble Subspace discriminant.

Then, the proposed model is further validated based on the Receiver Operating Characteristic Curve (ROC), a graphical representation that illustrates the classification capacity of the system. This curve is constructed by plotting True Positive Rate (TRP) against False Positive Rate (FPR). Another metric derived from the ROC plot is the Area Under the Curve (AUC), which measures the model's overall quality. Fig. 5 shows the ROC plot with the AUC of at most 1 for all the TIM plants.

B. Comparison With Other Medicinal Plant Classification Models

This section presents the performance comparison of the proposed model with other state-of-the-art techniques for medicinal plant classification. As the proposed model is trained and validated with the VNPlant-200 dataset, the training and validation performance of the proposed model is compared with theother models based on the same Mendeley dataset. The proposed model is also based on Inception network but with a residual connections-based deep transfer learning network. Incorporating the residual connection is better for training performance and validation performance. Table IV portrays the classification performance with machine learning algorithms. Table V shows the Classification Performance with Discriminant Analysis.

Table VI shows that the proposed model reaches the maximum training accuracy of 97.5 in the 20th iteration and the equivalent training time and prediction speed of the proposed model–D2SE. Table VII shows the results of testing samples with the label and its accuracy based on the number of learners and subspace dimension. Researchers have recently proposed several other methods to identify medicinal plants from leaf images. The proposed model is compared with all the existing deep learning-based classification systems for medicinal with the same data set. Table VIII shows that the proposed deep transfer learning model provides the highest validation accuracy of 97.5% compared to all other models. It shows that the proposed model outperforms the other existing state-of-the-art deep learning models for identifying medicinal plants. Moreover, the proposed model D2SE is designed as Mobile application.



(a) Class 1 - ROC

(b) Class 2 - ROC

(c) Class 3 - ROC

(d) Class 4 - ROC

(e) Class 5 - ROC

(f) Class 6 - ROC

Fig. 5. D²SE- RoC plot.

TABLE IV. Dense Features – Classification Performance With Machine Learning Models

Sl. No.	Classifier	Total Misclassification Cost	Training time (Secs)	Prediction speed	Accuracy
1.	Linear SVM	201	393.78	76/sec	94.4
2.	Quadratic SVM	173	2334.8	20/sec	95.2
3.	Cubic SVM	170	670.98	33/sec	95.3
4.	Fine Gaussian SVM	3123	889.36	23/sec	13.3
5.	Medium Gaussian SVM	340	664.8	30/sec	90.6
6.	Coarse Gaussian SVM	367	682.75	29/sec	89.8
7.	Fine KNN	189	109.33	250/sec	94.8
8.	Medium KNN	326	63.083	240/sec	90.9
9.	Coarse KNN	792	64.776	240/sec	78.0
10.	Cosine KNN	288	60.143	270/sec	92.0
11.	Cubic KNN	337	1869.4	8/sec	90.6
12.	Weighted KNN	265	68.86	260/sec	92.6
13.	Gaussian Naïve Bayes	507	143.82	1000/sec	85.9
14.	Kernel Naïve Bayes	592	61979	3/sec	93.6
15	**Liner Discriminant Analysis**	**166**	**118.45**	**160/sec**	**95.4**

TABLE V. Dense Features – Classification Performance With Discriminant Analysis

Sl. No.	Discriminant	Covariance structure	Training time	Prediction speed	Accuracy
1.	Linear Discriminant Analysis	Full	118.45	160/sec	95.4
2.	Linear Discriminant Analysis	Diagonal	35.893	1400/sec	83.5
3.	Quadratic Discriminant Analysis	Full	NA	NA	NA
4.	Quadratic Discriminant Analysis	Diagonal	15.628	1300/sec	85.9
5.	Optimized Discriminant	Full	1765.3	200/sec	95.4
6.	**Ensemble Subspace Discriminant**	**Full**	**872.65**	**21/sec**	**97.5**

TABLE VI. Dense Features – Classification Performance With Discriminant Analysis Based on Training Time and Prediction Speed

Sl. No.	Discriminant	Training time	Prediction speed	Accuracy
1.	Ensemble Boosted Trees	2144.2	3500/sec	48.4
2.	**Ensemble Bagged Trees**	**87.521**	**3500/sec**	**89.3**
3.	**Ensemble Subspace Discriminant**	**832.26**	**22/sec**	**97.5**

TABLE VII. Dense Features – Classification Performance With Ensemble Subspace Discriminant

Sl. No.	Discriminant	No. of learners	Subspace Dimension	Learner type	Accuracy
1.	Ensemble Subspace Discriminant	30	960	Nearest Neighbors	91.8
2.	**Ensemble Subspace Discriminant**	**30**	**960**	**Discriminant**	**97.5**

TABLE VIII. Performance Comparison of the Proposed Model With Other State-of-the-art Models

Authors	Accuracy
(Almazaydeh, 2022) [38]	95.7%
D²SE (Proposed Model)	**97.5%**

V. Conclusion

This paper proposes a robust deep learning app to identify the diabetic plants from VNPlant200 data set using fused deep neural network (DNN) model with Discriminant Subspace Ensemble. First, the features are extracted using DenseNet20 and the LDA is used to reduce the redundant information in this feature matrix and improve the discriminant ability of feature representation, and a nearest neighbors technique is used to produce a subspace ensemble for final diabetic therapeutic medicinal plant image classification. The proposed D²SEis deployed as a mobile application. The experiment results indicate that the proposed model achieved a reasonable recognition rate and provide a theoretical framework for further research and development of a medicinal plant classification system and positively contribute to the sustainability of human health. Our future research will focus on the study of the classification of medicinal plant leaves based on the leaf vein features and edge features.

Data availability: This work is based on VNPlant-200 dataset. It is publicly available in:

https://github.com/kencoca/VietNam-Medicinal-Plant

References

[1] International diabetic federation, “Facts and figures.” Accessed: Nov. 18, 2023. [Online]. Available at: www.idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html.

[2] K. Schreck and M. F. Melzig, “Traditionally used plants in the treatment of diabetes mellitus: screening for uptake inhibition of glucose and fructose in the Caco2-cell model,” *Frontiers in Pharmacology*, vol. 12, pp. 1-12, 2021.

[3] I. Süntar, “Importance of ethnopharmacological studies in drug discovery: role of medicinal plants,” *Phytochemistry Reviews*, vol. 19, no. 5, pp. 1199-1209, 2020.

[4] F. Majolo, L. K. de Oliveira Becker Delwing, D. J. Marmitt, I. C. Bustamante-Filho, and M. I. Goettert, “Medicinal plants and bioactive natural compounds for cancer treatment: important advances for drug discovery,” *Phytochemistry Letters*, vol. 31, pp. 196-207, 2018.

[5] K. Ahmad, I. Choi, and Y. H. Lee, “Implications of skeletal muscle extracellular matrix remodeling in metabolic disorders: diabetes perspective,” *International Journal of Molecular Sciences*, vol. 21, no. 11, pp. 3845, 2020.

[6] M. Mathew and S. Subramanian, “In vitro evaluation of anti-Alzheimer effects of dry ginger (Zingiber officinale Roscoe) extract,” *Indian Journal of Experimental Biology*, vol. 52, no. 6, pp. 606-612, 2014.

[7] J. Martins and S. Brijesh, “Phytochemistry and pharmacology of anti-depressant medicinal plants: a review,” *Biomedicine & Pharmacotherapy*, vol. 104, pp. 343-365, 2018.

[8] M. Setorki, “Medicinal herbs with anti-depressant effects,” *Journal of Herbmed Pharmacology*, vol. 9, no. 4, pp. 309-317, 2020.

[9] S. Sekhon-Loodu and H. P. V. Rupasinghe, “Evaluation of antioxidant, antidiabetic and antiobesity potential of selected traditional medicinal plants,” *Frontiers in Nutrition*, vol. 6, pp. 1-11, 2019.

[10] C. I. Chukwuma, M. G. Matsabisa, M. A. Ibrahim, O. L. Erukainure, M. H. Chabalala, and M. S. Islam, “Medicinal plants with concomitant anti-diabetic and anti-hypertensive effects as potential sources of dual acting therapies against diabetes and hypertension: a review,” *Journal of Ethnopharmacology*, vol. 235, pp. 329-360, 2018.

[11] Z. Rabiei, K. Solati, and H. Amini-Khoei, “Phytotherapy in treatment of Parkinson’s disease: a review,” *Pharmaceutical Biology*, vol. 57, no. 1, pp. 355-362, 2019.

[12] B. Joseph and D. Jini, “Antidiabetic effects of Momordica charantia (bitter melon) and its medicinal potency,” *Asian Pacific Journal of Tropical Disease*, vol. 3, no. 2, pp. 93-102, 2013.

[13] M. Son, Y. Miura, and K. Yagasaki, “Mechanisms for antidiabetic effect of gingerol in cultured cells and obese diabetic model mice,” *Cytotechnology*, vol. 67, no. 4, pp. 641-652, 2015.

[14] B. Y. Lai, T. Y. Chen, S. H. Huang, T. F. Kuo, T. H. Chang, C. K. Yang, M. T. Yang, and C. L. T. Chang, “Bidens pilosa formulation improves blood homeostasis and β-cell function in men: a pilot study,” *Evidence-Based Complementary and Alternative Medicine*, p. 832314, 2015.

[15] Z. Rabiei, K. Solati, and H. Amini-Khoei, “Phytotherapy in treatment of Parkinson’s disease: a review,” *Pharmaceutical Biology*, vol. 57, no. 1, pp. 355-362, 2019.

[16] C. J. Bailey, “The current drug treatment landscape for diabetes and perspectives for the future,” *Clinical Pharmacology & Therapeutics*, vol. 98, no. 2, pp. 170-184, 2015.

[17] F. Farzaei, M. R. Morovati, F. Farjadmand, and M. H. Farzaei, “A mechanistic review on medicinal plants used for diabetes mellitus in traditional persian medicine,” *Journal of Evidence-Based Integrative Medicine*, vol. 22, no. 4, pp. 944-955, 2017.

[18] A. B. Shori, “Screening of antidiabetic and antioxidant activities of medicinal plants,” *Journal of Integrative Medicine*, vol. 13, no. 5, pp. 297-307, 2015.

[19] M. Fatima, S. Hussain, M. Babar, M. Shahzad, and M.S. Zafar, "Curcumin: A Potential Therapeutic Agent for Human Health and Diseases–Current Status and Future Perspectives," *Pharmacological Benefits of Natural Agents*, pp. 278-297, 2023.

[20] J. M. Isaza, S. M. Gutiérrez, and D. R. S. Escobar, “Phytotherapy in Mexico: a review,” *Journal of Ethnopharmacology*, vol. 300, p. 113872, 2020.

[21] R. Gupta and B. Gigras, “Medicinal plants and natural products in amelioration of arsenic toxicity: a short review,” *Phytochemistry Reviews*, vol. 9, no. 2, pp. 23-33, 2010.

[22] T. Galochkina, M. Ng Fuk Chong, L. Challali, S. Abbar, and C. Etchebest, “New insights into GluT1 mechanics during glucose transfer,” *Scientific Reports*, vol. 9, no. 1, pp. 1-14, 2019.

[23] M. Kuroda, Y. Mimaki, Y. Sashida, T. Mae, H. Kishida, T. Nishiyama, M. Tsukagawa, E. Konishi, K. Takahashi, T. Kawada, K. Nakagawa, and M. Kitahara, “Phenolics with PPAR-γ ligand-Binding activity obtained from licorice (Glycyrrhiza uralensis Roots) and ameliorative effects of glycyrin

on genetically diabetic KK-Ay mice," *Bioorganic & Medicinal Chemistry Letters*, vol. 13, no. 24, pp. 4267-4272, 2003.

[24] F. Vlavcheski, D. Baron, I. A. Vlachogiannis, R. E. K. Macpherson, and E. Tsiani, "Carnosol increases skeletal muscle cell glucose uptake via AMPK-dependent GLUT4 glucose transporter translocation," *International Journal of Molecular Sciences*, vol. 19, no. 5, p. 1321, 2018.

[25] G. P. Kamatou, C. Ssemakalu, and L. J. Shai, "Cassia abbreviata enhances glucose uptake and glucose transporter 4 translocation in C2C12 mouse skeletal muscle cells," *Journal of Evidence-Based Integrative Medicine*, vol. 26, pp. 1-11, 2021.

[26] A. Altemimi, N. Lakhssassi, A. Baharlouei, D. G. Watson, and D. A. Lightfoot, "Phytochemicals: Extraction, isolation, and identification of bioactive compounds from plant extracts," *Plants*, vol. 6, no. 4, p. 42, 2017.

[27] T. N. Quoc and V. T. Hoang, "Medicinal Plant identification in the wild by using CNN," in *2020 International Conference on Information and Communication Technology Convergence (ICTC)*, 2020, pp. 25-29.

[28] C. Amuthalingeswaran, M. Sivakumar, P. Renuga, S. Alexpandi, J. Elamathi, and S. S. Hari, "Identification of medicinal plants and their usage by using deep learning," in *2019 3rd International conference on trends in electronics and informatics (ICOEI)*, 2019, pp. 886-890.

[29] Y. A. Putri, E. C. Djamal, and R. Ilyas, "Identification of medicinal plant leaves using convolutional neural network," in *Journal of Physics: Conference Series*, vol. 1845, no. 1, p. 012026, 2021.

[30] M. R. Dileep and P. N. Pournami, "Ayurveda: a deep learning approach for classification of medicinal plants," in *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)*, 2019, pp. 321-325.

[31] Y. Liu, Y. Li, Y. Zhao, and X. Na, "Image Classification and Recognition of Medicinal Plants Based on Convolutional Neural Network," in *2021 IEEE 21st International Conference on Communication Technology (ICCT)*, 2021, pp. 1128-1133.

[32] S. Liu, W. Chen, and X. Dong, "Automatic classification of Chinese herbal based on deep learning method," in *2018 14th International conference on natural computation, fuzzy systems, and knowledge discovery (ICNC-FSKD)*, 2018, pp. 235-238.

[33] A. Muneer and S. M. Fati, "Efficient and automated herbs classification approach based on shape and texture features using deep learning," *IEEE Access*, vol. 8, pp. 196747-196764, 2020.

[34] G. Mukherjee, B. Tudu, and A. Chatterjee, "A convolutional neural network-driven computer vision system for identifying species and maturity stage of medicinal leaves: case studies with Neem, Tulsi, and Kalmegh leaves," *Soft Computing*, vol. 25, no. 22, pp. 14119-14138, 2021.

[35] S. H. Lee, C. S. Chan, and P. Remagnino, "Multiorgan plant classification based on convolutional and recurrent neural networks," *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4287-4301, 2018.

[36] V. Tiwari, R. C. Joshi, and M. K. Dutta, "Deep neural network for multi-class classification of medicinal plant leaves," *Expert Systems*, e13041, 2022.

[37] M. Bhuiyan, M. Abdullahil-Oaphy, R. S. Khanam, and M. Islam, "MediNET: A deep learning approach to recognize Bangladeshi ordinary medicinal plants using CNN," in *Soft Computing Techniques and Applications*, 2021, pp. 371-380.

[38] L. Almazaydeh, R. Alsalameen, and K. Elleithy, "Herbal leaf recognition using mask-region convolutional neural network (mask r-CNN)," *Journal of Theoretical and Applied Information Technology*, vol. 100, no. 11.

[39] Kirtikar, Basu, "Indian Medicinal Plants (4 Vols. Set)," *Periodical Experts Book Agency*, 2012, ISBN-13: 9788171360536.

[40] NISC, "The Wealth of India: A Dictionary of Indian Raw Materials &Industrial Products," *National Institute of Science Communication, Council of Scientific & Industrial Research*, 2004, ISBN-13: 9788172362089.

[41] Sivarajan, Indira Balachandran, "Ayurvedic Drugs and Their Plant Sources," *CBS Publishers & Distributors Pvt Ltd, India*, 2017, ISBN-13: 9788120408289.

[42] Asolkar, Kakkar and Chakre, "Second Supplement to Glossary of Indian Medicinal Plants with Active Principles (Part - 1, A to K)," *National Institute of Science Communication*, 2010, ISBN-13: 9788120408289.

[43] Chopra; Nayar; Chopra, "Glossary of Indian Medicinal Plants," *The Publications & Information Directorate*, CSIR, 2022, ISBN-13: 9788172361266.

[44] Khare, C.P, "Indian Medicinal Plants: An Illustrated Dictionary," *Springer*, 2008, ISBN-13: 9780387706375.

[45] CSIR, "The Treatise on Indian Medicinal Plants," *The Publications & Information Directorate, CSIR*, 2010, ISBN-13: 9788172363130.

[46] Warrier, Nambiar, Ramankutty, "Indian Medicinal Plants: A Compendium of 500 Species (Vol. V)," *Orient Black Swan*, 2022, ISBN-13: 9788173717062.

[47] C. H. Chen and K. W. Huang, "Digit Recognition Using Composite Features With Decision Tree Strategy," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 2, pp. 98-107, 2023. DOI: 10.9781/ijimai.2022.12.001.

[48] V. Rajinikanth, S. Kadry, and P. Moreno-Ger, "ResNet18 Supported Inspection of Tuberculosis in Chest Radiographs With Integrated Deep, LBP, and DWT Features," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 2, pp. 38-46, 2023. DOI: 10.9781/ijimai.2023.05.004.

[49] Z. H. Arif, M. Mahmoud, K. H. Abdulkareem, S. Kadry, M. A. Mohammed, M. N. Al-Mhiqani, A. S. Al-Waisy, and J. Nedoma, "Adaptive Deep Learning Detection Model for Multi-Foggy Images," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 7, pp. 26-37, 2022. DOI: 10.9781/ijimai.2022.11.008.

N. Sasikaladevi

Dr. N. Sasikaladevi received Ph. D. Degree in Computer Science in 2013. She has published more than 42 papers in reputed international journals and conferences including publications in SCI-Indexed Journals. She is a reviewer of more than a dozen of reputed journals including IEEE transactions on Services Computing and IEEE Journal of Internet of Things. Her significant research contribution includes the study of genus-2 and genus-3 Hyper elliptic curve to design the security system for power constraints devices and applications. Her research focus includes the design of machine learning strategies to solve Discrete Logarithm Problem. She published several books and chapters in reputed publisher including Prentice Hall of India, Lambert Academic Publisher, IGI Global, Springer and Science Direct. She has received fund from Department of Science and Technology, Government of India to carry out projects in the domain of Services computing and Security System. She has received young scientist award and women scientist award from DST, India. She is also involved in the security enriched payment system design as a part of Digital India Initiative.

A. Revathi

Dr. A. Revathi has obtained B.E (ECE), M.E (Communication Systems), and Ph. D (Speech Processing) from National Institute of Technology, Tiruchirappalli, Tamil Nadu, India in 1988, 1993 and 2009 respectively. She has been serving on the faculty of Electronics and Communication Engineering for 34 years and she is currently working as a Professor in the Department of ECE, SASTRA Deemed University, Thanjavur, India. She has published 60 papers in Reputed International journals and presented papers in more than 80 International Conferences. Her areas of interest include Speech processing, Signal processing, Image processing, Biometrics and Security, Communication Systems, Embedded Systems and Computer Networks.

Pradeepa Sampath

Dr. Pradeepa Sampath received B. Tech degree in Information Technology from Jayaram College of Engineering, Tirucirappalli, India in 2005. She obtained M.Tech. in computer science from PSNA College of Engineering and Technology, Tiruchirappalli in 2012. She received Ph. D. Degree in Computer Science in 2013 in the Department of Computer Science and Engineering, School of Computing, SASTRA Deemed University, India. She has published 15 papers in Reputed International journals and presented papers in more than 8 International Conferences. Her areas of interest include Text Mining, Machine Learning, Deep Learning, Big Data Analytics.

Vimal Shanmuganathan

Dr Vimal Shanmuganathan has obtained Ph.D degree in 2019 and completed the Post Doc in 2023. He is working in the Department of Artificial Intelligence & Data Science, Sri Eshwar College of Engineering, Coimbatore, Tamil Nadu (India). He has around Fifteen years of teaching experience, EMC certified Data science Associate and CCNA Certified professional too. He holds a Ph.D in Information and Communication Engineering from Anna University Chennai and he received Masters Degree from Anna University Coimbatore. His areas of interest include Game Modelling, Artificial Intelligence, Cognitive radio networks, Network security, Machine Learning and Big data Analytics. He is a Senior member in IEEE and holds membership in various professional bodies. He has hosted 22 special issues in IEE, Elsevier, Springer, Wiley and CMC tech science journals. He has served as Guest editor for SCI journals and editored 3 books in Scopus indexed also hosted 3 International conference indexed by Scopus. He is a Senior Member in IEEE.

Gaurav Dhiman

Dr. Gaurav Dhiman is a highly accomplished academician and researcher with an impressive track record of excellence in computer science. He holds a Ph.D. in Computer Engineering from Thapar Institute of Engineering & Technology, Patiala, and has also completed his Master Degree of Computer Applications from the same institute. Currently, he is serving as an Assistant Professor in the School of Sciences and Emerging Technologies, Jagat Guru Nanak Dev Punjab State Open University, Patiala. His research has been recognized at the highest level, as he has been named one of the world's top researchers by Stanford University, USA list of world's top 2% scientists prepared by Elsevier. He is the senior member of IEEE. He has also received accolades for his outstanding reviewer work from Knowledge-Based Systems (Elsevier). He has authored over 300 peer-reviewed research papers, all of which are indexed in SCI-SCIE, and 10 international books. He is currently serving as a guest editor for more than forty special issues in various peer-reviewed journals. He is an Editor-in-Chief of the International Journal of Modern Research (IJMORE), Co-Editor-in-Chief of the International Journal of Electronics and Communications Systems (IJECS) and International Journal of Ubiquitous Technology and Management (IJUTM). He is an Associate Editor of IEEE Transactions on Industrial Informatics, IET Software (Wiley), Expert Systems (Wiley), IEEE Systems, Man, and Cybernetics Magazine, Spatial Information Research (Springer), and more. He is also an Academic Editor of Computational and Mathematical Methods in Medicine (Hindawi), Scientific Programming (Hindawi), Mobile Information Systems (Hindawi), and Review Editor of Frontiers in Energy Research - Fuel Cells, Frontiers in High-Performance Computing - Parallel and Distributed Software, Frontiers in Artificial Intelligence - Pattern Recognition, and more. His research interests span a wide range of topics including Artificial Intelligence, Internet of Things, Machine-learning, Deep-learning, Single-, Multi-, Many-objective optimization (Bio-inspired, Evolutionary, and Quantum), Soft computing (Type-1 and Type-2 fuzzy sets), Power systems, and Change detection using remotely sensed high-resolution satellite data. He has published his research in various reputed journals, including IEEE, Elsevier, Springer, Wiley, Taylor and Francis, among others.

Efficient Gated Convolutional Recurrent Neural Networks for Real-Time Speech Enhancement

Fazal-E-Wahab[1]*, Zhongfu Ye[1], Nasir Saleem[2], Hamza Ali[3], Imad Ali[4]

[1] National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China, Hefei, 230026 Anhui (China)
[2] Department of Electrical Engineering, Faculty of Engineering & Technology, Gomal University, D.I. Khan, KPK (Pakistan)
[3] Department of Electrical Engineering, University of Engineering & Technology, Mardan, KPK (Pakistan)
[4] Department of Computer Science, University of Swat, KPK (Pakistan)

* Corresponding author: wmarwat@mail.ustc.edu.cn

Received 5 March 2022 | Accepted 27 December 2022 | Published 31 May 2023

Abstract

Deep learning (DL) networks have grown into powerful alternatives for speech enhancement and have achieved excellent results by improving speech quality, intelligibility, and background noise suppression. Due to high computational load, most of the DL models for speech enhancement are difficult to implement for real-time processing. It is challenging to formulate resource efficient and compact networks. In order to address this problem, we propose a resource efficient convolutional recurrent network to learn the complex ratio mask for real-time speech enhancement. Convolutional encoder-decoder and gated recurrent units (GRUs) are integrated into the Convolutional recurrent network architecture, thereby formulating a causal system appropriate for real-time speech processing. Parallel GRU grouping and efficient skipped connection techniques are engaged to achieve a compact network. In the proposed network, the causal encoder-decoder is composed of five convolutional (Conv2D) and deconvolutional (Deconv2D) layers. Leaky linear rectified unit (ReLU) is applied to all layers apart from the output layer where softplus activation to confine the network output to positive is utilized. Furthermore, batch normalization is adopted after every convolution (or deconvolution) and prior to activation. In the proposed network, different noise types and speakers can be used in training and testing. With the LibriSpeech dataset, the experiments show that the proposed real-time approach leads to improved objective perceptual quality and intelligibility with much fewer trainable parameters than existing LSTM and GRU models. The proposed model obtained an average of 83.53% STOI scores and 2.52 PESQ scores, respectively. The quality and intelligibility are improved by 31.61% and 17.18% respectively over noisy speech.

Keywords

Convolutional Recurrent Networks, Deep Learning, GRU, Intelligibility, LSTM, Speech Enhancement.

DOI: 10.9781/ijimai.2023.05.007

I. Introduction

Speech enhancement (SE) aims to suppress background noise signals from the target speech, which include non-speech noise, competing speech, and room reverberations [1]. SE is used as a front-end in various real-world applications such as robust ASR systems and mobile phone communications where real-time processing is required. In such applications, SE is required to perform with little computational complexity and provide near-instantaneous outputs. The aim of this study is to focus on single-microphone speech enhancement, operating in real-time systems. For listeners using digital hearing aids, a delay of 3 milliseconds is perceptible, whereas delays longer than 10 msec are intolerable [2]. Speech enhancement techniques have made significant progress during the last several decades. Speech enhancement techniques may be divided into two categories depending on the quantity of microphones used, that is, single-channel based and multi-channel. The high-availability and low-cost single-channel approaches nevertheless have significant research significance, even if the extra spatial information of the microphone array may assist in reducing the direction-related noise interference. As a result, the goal of this work is to concentrate on real-time, single-microphone speech enhancement. Delays of 3 milliseconds or less are noticeable to listeners, whereas those of 10 milliseconds or more are unpleasant [2]. In these situations, a causal SE system is often necessary to prevent delays. A Causal SE system is often a requisite in such applications to avoid delays.

In recent years, SE has been formulated as a supervised learning problem where a deep neural network (DNN) learns a mapping

Please cite this article as:
F. E. Wahab, Z. Ye, N. Saleem, H. Ali, I. Ali. Efficient Gated Convolutional Recurrent Neural Networks for Real-Time Speech Enhancement, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 1, pp. 66-74, 2024, http://dx.doi.org/10.9781/ijimai.2023.05.007

function from the noisy features to a time-frequency mask [3]. The ideal binary mask, which categorises time-frequency units into the speech-dominant and noise-dominant, was the first training-target used in supervised speech enhancement [3]. More recent training-targets include ideal ratio mask (IRM) [4]-[5] complex ratio mask [6], mapping-based targets related to the magnitude or power spectra of the target speech [7]-[8], ideal amplitude mask [9]. For supervised SE, both noise and speaker generalizations are vital. An easy but useful approach to cope with noise generalization is to train a network with large noise types [10]. Likewise, a large number of speakers can be used in a training set to deal with speaker generalization. However, in the presence of several training speakers, a feedforward DNN is inept at tracking a target-speaker [10]-[11].

The training-targets in the time-frequency domain are mainly divided into two classes: the masking-based and mapping-based targets. The masking-based targets describe a time-frequency relationship between the clean speech and background noise signals, whereas the mapping-based targets correspond to the spectral representations of clean speech. In the masking-based class, ideal binary mask (IBM) [12], ideal ratio mask (IRM) [4] and spectral magnitude mask (SMM) [4] only use the magnitude between clean speech and mixture speech, overlooking the phase spectrum. Alternatively, the phase-sensitive mask (PSM) [13] incorporated the phase information and showed the importance of phase spectrum estimation. Afterward, complex ratio mask (cRM) [6] can be used to recover speech efficiently by improving both real and imaginary parts of the clean and noisy speech spectrograms simultaneously. Recently, [14] proposed a convolutional recurrent network using one encoder and two decoders to estimate the real and imaginary spectrograms (complex spectral mapping) of the noisy speech concurrently. It is important that the complex ratio mask and complex spectral mapping obtain the complete information of a speech signal to accomplish the best SE performance. The convolutional encoder-decoder (CED) and GRU produce a convolutional recurrent neural network (CRN), which is used to develop the SE in this article. CED is a strong tool for extracting temporal and spatial patterns from raw data. A causal system that is suitable for real-time speech processing is created by integrating a convolutional encoder-decoder and GRUs into the convolutional recurrent network architecture. In comparison to typical RNNs, GRU has the capacity to learn long-term temporal dependencies in speech signals with a far smaller number of trainable parameters. The contributions of this study are summarized as:

- A causal SE system which is appropriate for real-time speech processing is created by integrating a convolutional encode-decoder and GRUs into the convolutional recurrent network architecture.
- For an appropriate shape of the inputs required by GRUs, the proposed model has grouped the fully connected recurrent neural networks into disconnected parallel recurrent neural networks, where the forward information flow remains the same.
- By adding the skipped connections, to avoid gradient decay, which connect the output of the encoder to input of decoder output doubles the feature Maps, results in increasing the model complexity. Therefore, in the proposed model, add-skipped connection between conv-deconv layers having (1×1) kernel size is proposed which improves network performance at negligible complexity.

The remaining of the paper is organized as follows. Related studies on the research are presented in Section II. Description of the proposed model is given in Section III. The experimental setup and results are given in Section IV. Finally, the paper is concluded in Section V.

II. Related Studies

Generally, a DNN individually predicts labels for all time frames using small context windows and cannot control long-term context windows that are essential for target speaker tracking. Recently, studies [11], [15] suggest that it is better to formulate SE as a sequence-to-sequence process to control the long-term context windows. Recurrent Neural Networks [16] and Convolutional Neural Networks have been employed with such a formulation where training and testing with different noise types and speakers can be carried out. A four-layer LSTM model for speaker generalization is proposed in [15]. The results showed that the LSTM model generalized better to untrained speakers and considerably outperformed a DNN-based model in terms of speech intelligibility. Recently, a dilated convolution-based gated residual network is developed in [17]-[18] which demonstrated better generalization potential for untrained speakers at various SNRs when compared to the LSTM by [10]. However, the gated residual network requires future information for spectral masking or spectral mapping. Thus, it is not suitable for real-time SE. Motivated by recent studies [19]-[20] on convolutional recurrent networks; we designed a compact and efficient architecture for real-time speech enhancement. The first convolutional encode-decoder architecture has been introduced for SE by [21]. A redundant convolutional encode-decoder [22] was proposed, based on the convolution repetitions, batch normalization, and a ReLU activation layer. Moreover, to facilitate network optimization, skip connections are used. In this study, a skips-based convolutional encoder-decoder and the parallel GRUs are integrated into a convolutional recurrent architecture to estimate the complex ratio mask (cRM). We observed that the proposed architecture provided improved speech quality and intelligibility as compared to the GRU and LSTM with fewer trainable parameters. A deep residual GRU-based model to enhance noisy speech was proposed [23] which performed better as compared to SOTA for speech enhancement and recognition tasks. The study in [24] presented a joint structure to solve single-channel speech enhancement in the complex-domain. The RBM in [25] is extended for spectral masking and noisy speech enhancement. The acoustic features in traditional RBMs are extracted layerwise, where feature compression results in a loss of information during training. To address this problem of retaining information in raw speech, RBMs are extended for acoustic feature representation and speech enhancement. Acoustic features and regularized sparse features are combined to train DNNs for better speech enhancement [26]. Using short context windows, FNN model [27] independently predicts labels for all time frames. The CNN [28]-[29] may learn local features involved in the training data, in contrast to the FNN [8]-[9] which can fully use the previous knowledge of speech. The long-term contexts of speech signals cannot be leveraged by either the FNN or the CNN model, however. In order to regulate the long-term context windows, it has recently been recommended by the research [30]-[31], that it is preferable to design SE as a sequence-to-sequence process. The RNN model [32] can deal with long-term contexts in a sequence-based way, but often needs very complex hand-crafted features like MFCC. Convolutional recurrent networks (CRN) were first used for speech improvement by combining the CNN and RNN [31], [33]. Convolutional and recurrent neural networks [34] have been used in a formulation that enables training and testing with a variety of speakers and noise sources. Due to high computational load, most of the DL models for speech enhancement are difficult to implement for real-time processing. It is challenging to formulate resource efficient and compact networks.

III. Proposed System Description

The convolutional encoder-decoder (CED) and GRU produce a convolutional recurrent neural network (CRN), which is used to develop the SE in this article. CED is a strong tool for extracting temporal and spatial patterns from raw data. A causal system that is suitable for real-time speech processing is created by integrating a convolutional encoder-decoder and GRUs into the convolutional recurrent network architecture. In comparison to typical RNNs, GRU has the capacity to learn long-term temporal dependencies in speech signals with a far smaller number of trainable parameters. CED and GRU are explained in the following subsections.

A. Causal Convolution-Based Encoder-Decoder

The encoder in the causal convolution-based encoder-decoder framework is made up of stacked convolutional and pooling layers that extract high-level features from raw input data. Fundamentally similar structure as the encoder but in the reverse order, the decoder maps low-level features at the encoder output to full input feature size. This symmetric structure of the encoder-decoder ensures the shape of inputs and outputs. We imposed causal convolutions on the encoder-decoder framework to design a real-time SE system. Fig. 1. illustrates the causal convolutions with time-dimension. We treat the inputs as the sequence of the feature vectors, whereas the outputs are independent of the future sequence of the feature vectors. With such causal convolutions, the architecture leads to a causal encoder-decoder framework. The causal deconvolution can easily be applied to the decoder.

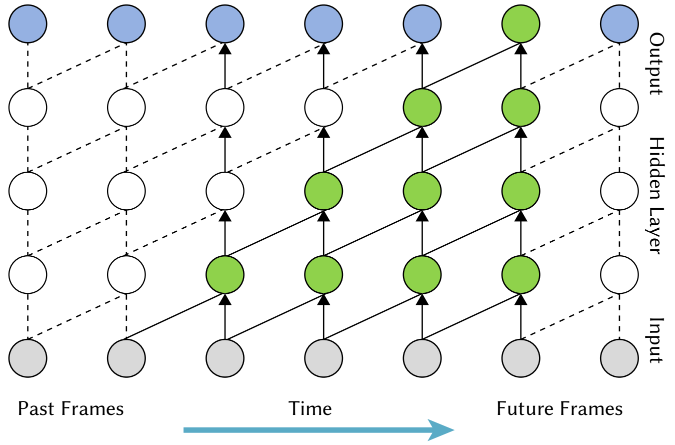


Fig. 1. An example of causal convolutions. The convolution output does not depend on future inputs.

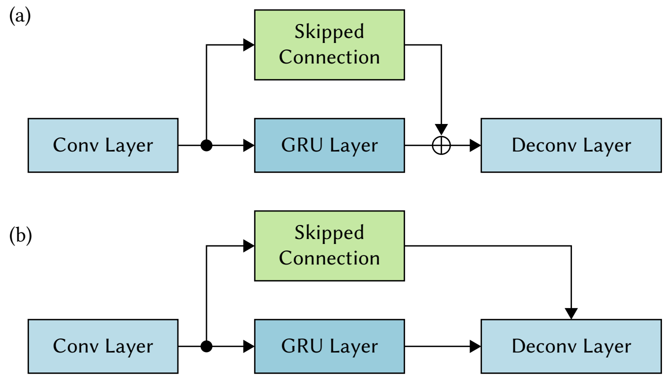


Fig. 2. Skip connections (a) add-skipped connections, (b) doubling the decoder inputs.

In the proposed network, the causal encoder-decoder is composed of five convolutional (Con2D) and deconvolutional (Decon2D) layers. Leaky linear rectified unit (ReLU) [35]-[36] is applied to all layers apart from the output layer where softplus activation (which can confine the network output to always be positive) [37] is utilized. Leaky ReLU has shown fast convergence and better generalization. Furthermore, batch normalization is adopted after every convolution (or deconvolution) and prior to activation. The kernel number is increased steadily in the encoder, whereas it is decreased steadily in the decoder, such that symmetric kernel numbers are adopted. So as to leverage large contexts, a stride of 2 is used along the frequency direction for all the convolutional (or deconvolutional) layers, whereas the time dimension of the features remains the same. To get a better flow of gradients and information all through the network, skip connections are utilized which connect the encoder outputs to the decoder inputs, as depicted in Fig. 2(a). In a recent study [17], the skip connections have been adopted by connecting the output of the encoder to the input of the decoder, as depicted in Fig. 2(b), which doubles the number of input channels to the decoder, resulting in increasing the complexity.

B. Temporal Modeling Using Parallel GRUs

Leveraging the long context is important to track a target speaker. The GRU [26] is the newer type of recurrent neural network that includes memory cells and is successful in temporal modeling. To integrate the temporal dynamics of the speech signals, we inserted a parallel GRU layer between the convolutional-encoder and the convolutional-decoder. Equations (1)-(4) describe the GRU network.

$$z_l = \sigma(W_z[x_t, h_{t-1}] + b_z) \tag{1}$$

$$r_t = \sigma(W_r[x_t, h_{t-1}] + b_r) \tag{2}$$

$$\tilde{h}_l = \tanh\,(W_h[r_l \odot h_{l-1}, x_l] + b_h) \tag{3}$$

$$h_l = (I - z_l) \odot h_{l-1} + (z_l) \odot \bar{h}_t \tag{4}$$

Where $z_t, r_t, \tilde{h}_l, h_l$ represent update gate, reset gate, intermediate memory, and output respectively whereas W_z, W_r, W_h, b_z, b_r and b_h are the model parameters that are learned during training. For a suitable shape of the inputs required by GRUs, the approach proposed by [17] has been adopted to group the wide fully connected recurrent neural networks into P disengaged parallel recurrent neural networks. But, noted that the forward information flow remains the same. The parallel GRUs are denoted by P, where P = 1 indicates that the last convolutional encoder output is flattened to a single vector and fed to a single GRU, whereas P > 1 indicates that the encoder output is reshaped to P vectors of the same length, fed through P disconnected GRUs, and reshaped again to the number of decoder channels. Another practical advantage is the possible parallel execution of the disconnected RNNs. It is important to note that the insertion of the GRUs does not impact the system's causality.

C. Network Architecture

In this paper, we used 161-dimensional STFT magnitude spectrum of noisy speech as the input features and a complex ratio mask as the training target. The proposed convolutional recurrent network is illustrated in Fig. 3, where inputs to the network are encoded into a high-dimension latent space, and the sequences of latent features are subsequently modelled by the GRU layer. Next, the output sequences of the GRU layer are transformed back by the decoder into their original input shape. The proposed convolutional recurrent network uses CNNs for feature extraction and RNNs for temporal modeling, thus combining two powerful topologies with improved results. A representation of the architecture is given in Table 1. The input and output layers' sizes are specified as FeatureMaps (FM), TimeSteps (TS), and FrequencyChannels (FCh), respectively, while

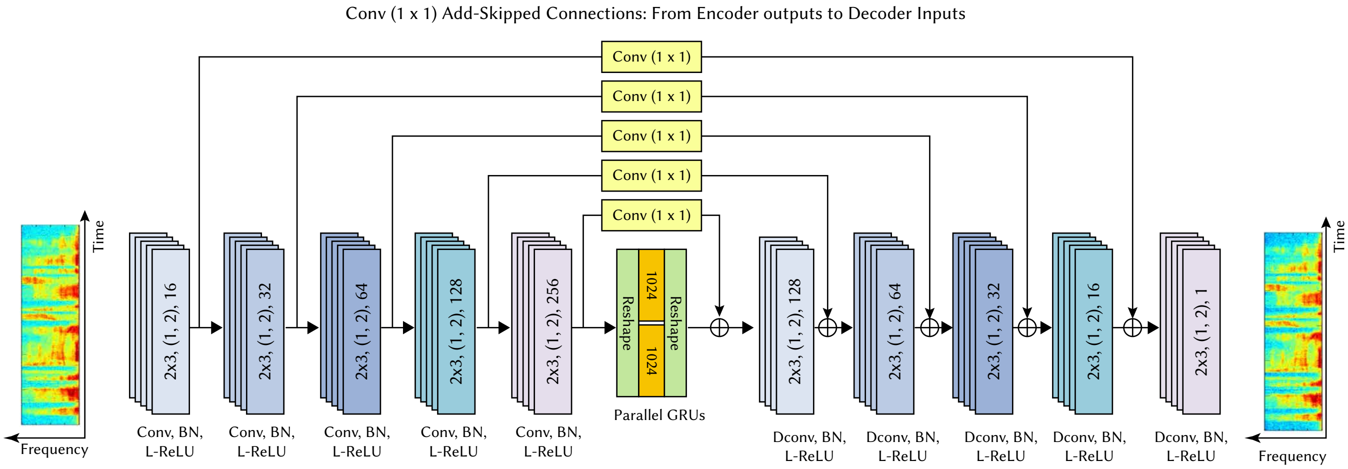


Fig. 3. Network architecture of our proposed CGCRN.

TABLE I. Network Architecture, Where *T* Denotes the Time Frames in the STFT Magnitude Spectrum, Here P = 2 in P-GRU Layer, Epochs= 100, and Learning Rate Is 0.0001

Layer	Input Size	Hyperparameters	Output Size
	FM × TS × FCh	KZ, S, OCh	FM × TS × FCh
Reshape-1	T × 161	---	1 × T × 161
Conv-1	1 × T × 161	2 × 3, (1, 2), 16	16 × T × 80
Conv-2	16 × T × 80	2 × 3, (1, 2), 32	32 × T × 39
Conv-3	32 × T × 39	2 × 3, (1, 2), 64	64 × T × 19
Conv-4	64 × T × 19	2 × 3, (1, 2), 128	128 × T × 9
Conv-5	128 × T × 9	2 × 3, (1, 2), 256	256 × T × 4
Reshape-2	256 × T × 4	---	T × 2048
P-GRU	T × 2048	2048	T × 1024
Reshape-3	T × 1024	---	256 × T × 4
Deconv-1	512 × T × 4	2 × 3, (1, 2), 128	128 × T × 9
Deconv-2	256 × T × 9	2 × 3, (1, 2), 64	64 × T × 19
Deconv-3	128 × T × 19	2 × 3, (1, 2), 32	32 × T × 39
Deconv-4	64 × T × 39	2 × 3, (1, 2), 16	16 × T × 80
Deconv-5	32 × T × 80	2 × 3, (1, 2), 1	1 × T × 4
Reshape-4	1 × T × 161	---	T × 161

the hyperparameters along the layer are specified as KernelSize (KZ), Strides (S), and OutChannels (OCh). In all the convolution and the deconvolution layers, a zero-padding in the time direction is applied, but no padding is involved in the frequency direction. For causal convolutions, a (2×3) kernel size is used, where (2×3) indicates (time×frequency). Note that by adding the skipped connections, which connect the output of the encoder to the input of the decoder output, doubles the feature maps, thus increasing the network complexity. By adding an add-skipped connection between the conv-deconv with (1×1) kernel size, it improves network performance at negligible added complexity, as shown in Fig. 4. We denoted the proposed network as CGCRN.

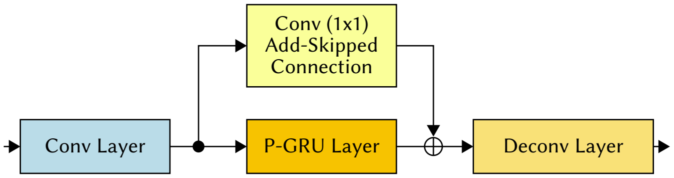


Fig. 4. (1 × 1) convolutions in the add-skipped connections.

D. LSTM and GRU Baselines

In the experiments, causal LSTM and GRU baselines were selected for comparison purposes. In the causal LSTM and GRU models, a context feature window of 11 frames, composed of 10 past speech frames and 1 current speech frame is used to estimate one frame of the target speech. A concatenated long vector of 11 frames of feature vectors is used as input to the network at all-time steps. We used the same network architectures for LSTM and GRU [11 161, 1024, 1024, 1024, 1024, 1024, 1024, 1024] units from the input to the output layer. No future information is used by baselines, which makes them causal speech enhancement systems.

IV. Experimental Setup

In the experiments, we evaluated SE networks on the LibriSpeech dataset [38] (derived from the read audiobooks, LibriVox project) including 0.25 Million utterances from 2.1k speakers. We have used the LibriClean version of LibriSpeech which contains 104014 clean utterances (about 360 hours) belonging to 921 different speakers. But to evaluate the networks used in this study, we randomly selected 5000 speech utterances from 40 speakers. Among these speakers, 2 male

TABLE II. Networks Comparison in Three Test-Noises in Terms of the STOI (In %)

Noise Types	Babble				Street				Cafeteria			
Input SNR	-5 dB	-2 dB	0 dB	Avg	-5 dB	-2 dB	0 dB	Avg	-5 dB	-2 dB	0 dB	Avg
Noisy Speech	58.95	66.30	75.55	66.93	58.30	66.20	75.08	66.52	57.40	65.19	74.21	65.60
LSTM	77.29	82.62	84.96	81.62	75.21	82.62	84.11	80.64	74.32	81.38	83.07	79.59
GRU	77.45	83.21	85.01	81.89	75.30	82.05	85.01	80.78	74.14	81.25	83.22	79.53
CRN	79.71	85.48	86.88	84.02	77.12	84.44	87.23	82.93	76.07	82.68	85.10	81.28
CNN-GRU	78.11	84.31	85.82	82.74	76.21	83.01	85.66	81.62	75.04	82.31	84.14	80.49
FCNN	70.22	75.21	80.34	75.26	71.44	75.57	81.02	76.00	70.34	76.70	80.87	75.97
CGCRN	**80.47**	**86.29**	**87.74**	**84.83**	**77.86**	**85.23**	**88.07**	**83.72**	**76.80**	**83.43**	**85.92**	**82.05**

TABLE III. Networks Comparison in Three Test-Noises in Terms of the PESQ

Noise Types	Babble				Street				Cafeteria			
Input SNR	-5 dB	-2 dB	0 dB	Avg	-5 dB	-2 dB	0 dB	Avg	-5 dB	-2 dB	0 dB	Avg
Noisy Speech	1.63	1.79	1.86	1.76	1.58	1.71	1.84	1.71	1.52	1.70	1.82	1.68
LSTM	2.06	2.36	2.53	2.32	2.03	2.31	2.48	2.27	2.04	2.30	2.47	2.27
GRU	2.07	2.36	2.54	2.32	2.05	2.27	2.47	2.26	2.03	2.31	2.48	2.27
CRN	2.17	2.44	2.62	2.41	2.14	2.40	2.60	2.38	2.12	2.38	2.59	2.36
CNN-GRU	1.95	2.26	2.53	2.25	1.98	2.31	2.55	2.28	1.94	2.25	2.50	2.23
FCNN	1.81	2.15	2.44	2.13	1.88	2.21	2.51	2.20	1.85	2.20	2.48	2.18
CGCRN	**2.29**	**2.59**	**2.79**	**2.56**	**2.25**	**2.53**	**2.76**	**2.52**	**2.21**	**2.49**	**2.72**	**2.47**

and 2 female speakers are used as untrained speakers, whereas the remaining 36 speakers are used to train the networks. In order to train noise-independent networks, we have used 60 noise types from the Perception and Neurodynamics Laboratory (http://web.cse.ohio-state.edu/pnl/data.html) and Laboratory for Recognition and Organization of Speech and Audio (https://www.ee.columbia.edu/~dpwe/sounds/) for network training. For testing purpose, we used three challenging noise types (multi-talker babble, street, and cafeteria). We created a training set by randomly selecting utterances with an indiscriminate cut from the 60 training noise types at SNRs selected from -5dB, -3dB, -1dB, 0dB, and 2dB. During testing, we used three SNRs for the test set, that is, -5dB, -2dB, and 2dB. In order to examine the speaker generalization, the models are tested with two test sets for all noise types (i.e., multi-talker babble, street, and cafeteria) using trained and untrained speakers, respectively. First test set is composed of 120 mixtures created from 30 × 4 utterances of 5 trained speakers, whereas the second test set is composed of 120 mixtures created from 30 × 4 utterances of 4 untrained speakers. Speech utterances and noise types are sampled at 16 kHz. The networks are optimized with the Adam optimizer [39]-[40]. We fixed the learning rate to 10-4 and the mean squared error (MSE) served as a loss function. The networks are trained with minibatch size of 16 and the number of epochs is fixed to 80. Inside all minibatches, the training samples are zero padded such that to contain the equal number of time steps.

The experiments use two widely used objective metrics to quantify the proposed speech enhancement, including the STOI (Short-Time Objective Intelligibility), the PESQ (Perceptual Evaluation of Speech Quality). Intelligibility and quality of the enhanced speech signals are determined by STOI and PESQ, respectively. PESQ [41], an ITU-T P.862 recommendation, scores the perceptual speech quality from -0.5 to 4.5. Similarly, STOI [42] assesses speech intelligibility with output values ranging from 0 to 100.

V. Results and Discussions

Two performance metrics are used in the experiments. Perceptual evaluation of speech quality (PESQ) [41] measures the speech quality whereas the short-term objective speech intelligibility (STOI) [42] measures the speech intelligibility, respectively. A high value for both the metrics indicates a better performance. We also included the Convolutional recurrent network (CRN) proposed by [17], CNN-GRU [43], and fully connected CNN (FCNN) [44] as SOTA for comparison. The proposed CGCRN network is the extension of CRN. Table II-III presents STOI and PESQ test scores of noisy speech and speech processed by different networks across all the noise types and input SNRs. The best performance is highlighted with boldface numbers. As indicated by Table II-III, the LSTM and GRU networks yielded almost similar STOI and PESQ scores which suggests that the noisy speech can effectively be enhanced by using GRU networks with less trainable parameters. The results indicated that replacing the LSTM layers by a parallel GRU layer in the CRN significantly improved the performance with fewer trainable parameters and network complexity. The CRN outperformed both the LSTM and GRU networks. On the other hand, the proposed CGCRN consistently outperformed the LSTM, GRU, and CRN in both the metrics. For example, the average STOI test scores are improved from 66.93% to 84.83% with CGCRN (ΔSTOI = 17.90%) in babble noise type. Here Δ indicates the improvements in metrics. Also, the average STOI test scores are improved from 65.60% to 82.05% with CGCRN (ΔSTOI = 16.45%) in cafeteria noise type. On average, 3.09% STOI gain is achieved when the CGCRN when compared to the LSTM network. Moreover, 0.79% improvement in STOI test scores is achieved against the CRN. In addition, the average PESQ test scores are improved from 1.76 to 2.56 with the proposed CGCRN (ΔPESQ = 0.80 equivalent to 31.25%) in babble noise type. Similarly, the average PESQ test scores are improved from 1.71 to 2.52 with the proposed CGCRN (ΔPESQ = 0.81 equivalent to 32.14%) in street noise type.

Table IV-V presents the speaker generalization potential of the neural networks used in this study. It can be observed from Tables that the CGCRN and CRN showed better generalization to untrained speakers. In the most challenging noisy cases, where the utterances from the untrained speakers are mixed with three untrained noise types at -5dB and -2dB, the proposed CGCRN improved the average STOI by 15.72% and the PESQ by 0.70 (28.29%) over the noisy speech. The CGCRN improved the PESQ by 9.16 % over the GRU in trained speakers whereas by 9.91% in untrained speakers, respectively. Thereby indicates that the proposed models can success-fully be implemented in untrained situations.

TABLE IV. Speaker Generalization of the Networks in Terms of STOI (In %)

Speaker Types	Trained Speakers				Untrained Speakers			
Noise Type	Babble	Street	Cafeteria	Avg	Babble	Street	Cafeteria	Avg
LSTM	81.62	80.64	79.59	80.61	79.32	78.22	77.01	78.18
GRU	81.89	80.78	79.53	80.73	79.57	78.41	77.19	78.39
CGCRN	**84.83**	**83.72**	**82.05**	**83.53**	**83.66**	**82.33**	**80.23**	**82.07**

TABLE V. Speaker Generalization of the Networks in Terms of PESQ

Speaker Types	Trained Speakers				Untrained Speakers			
Noise Types	Babble	Street	Cafeteria	Avg	Babble	Street	Cafeteria	Avg
LSTM	2.32	2.27	2.27	2.29	2.22	2.19	2.18	2.20
GRU	2.32	2.26	2.27	2.28	2.21	2.17	2.17	2.18
CGCRN	**2.56**	**2.51**	**2.47**	**2.51**	**2.45**	**2.43**	**2.38**	**2.42**

The batch normalization in convolution operations accelerated the training and improved the performance. We observed a faster convergence and less MSEs with the CGCRN as compared the LSTM and GRU networks. Importantly the fewer trainable parameters are the key significance of the CGCRN, as illustrated in Fig. 5. In addition, the causal convolution operations captured the local patterns in the magnitude spectra exclusive of the future in-formation. In contrast, the GRU and LSTM networks deal all input frames as flattened feature vectors, thereby lack ample control over the time-frequency structures in the magnitude spectra. The parallel GRUs layer models the long-term temporal dependencies in a compressed space which is vital to speaker classification in the speaker-independent SE. As shown in experiments, replacing the LSTM layers in CRN with a single parallel GRUs layer yielded a considerable performance gain and enormous computational savings. A single GRU layer reduces 25% of trainable parameters (network complexity) as com-pared to single LSTM layer.

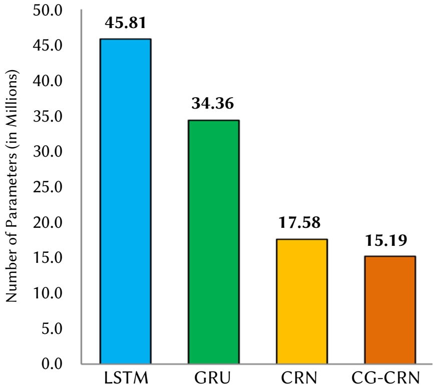


Fig. 5. Parameter efficiency comparison of different models. We compare the number of trainable parameters in different models.

Table VI shows impact of add-skipped connections in the CGCRN architecture. Adding the skipped connections is superior to no skipped connections. Although add-skips improved the PESQ and STOI test scores, but a better performance is achieved by inserting Conv (1 × 1) add-skipped connections. In order to visualize the spectrotemporal characteristics, the spectrograms are presented in Fig. 6. which belongs to the clean speech, noisy speech, and speech processed by the LSTM, GRU, and CGCRN with the cRM as the training-target. It is evident that few speech parts are missing in the spectrograms (highlighted with boxes) of speech enhanced by LSTM and GRU. In contrast, the speech enhanced by CGCRN demonstrates comparable spectrotemporal patterns to the clean speech and less distortion can also be noticed.

TABLE VI. Effects of Skipped Connections

Skip Types	STOI	PESQ
No Skips	79.21	2.34
Add Skips	81.33	2.40
Conv Skips	**83.45**	**2.49**

The proposed speech enhancement CGCRN performed better at all input SNRs in terms of speech intelligibility and quality. However, to confirm the success at SNRs, one-way analysis-of-variance (ANOVA) statistical analyses are conducted at -5dB, -2dB and 0dB. The statistical tests are performed at 95% confidence interval. Differences between test results are believed statistically important if the probability (Pvalue) is less than 0.05 (P<0.05) and Fvalue is higher than the critical value of FDistribution (Fvalue>FCritical). Table VII shows the statistical tests in terms of speech intelligibility at 95% confidence interval with FCritical is 3.09. It is clear that Pvalues of the proposed model are less than 0.05 and the values of FCritical are higher than 3.09, which indicates that the intelligibility results of the proposed model are statistically significant. Similarly, Table VIII shows the statistical tests in terms of speech intelligibility at 95% confidence interval with FCritical is 3.09. For illustration at adverse noise levels (-5dB), the CGCRN against the noisy speech (CGCRN ⟶ Noisy), we achieved [F (2, 100) = 39.5, p < 0.001] for STOI and [F (2, 100 = 32.2), p < 0.001] for PESQ, respectively. Also, against the CRN (CGCRN ⟶ CRN), we achieved [F (2, 100) = 21.1, p < 0.001] for STOI and [F (2, 100) = 24.1, p < 0.001] for PESQ. Moreover, against LSTM (CGCRN ⟶ LSTM), we achieved [F (2, 100) = 22.2, p < 0.001] for STOI and [F (2, 100) = 18.3, p < 0.015] for PESQ. The ANOVA results at low SNRs indicate that the proposed model achieved better results statistically over the competing deep learning models.

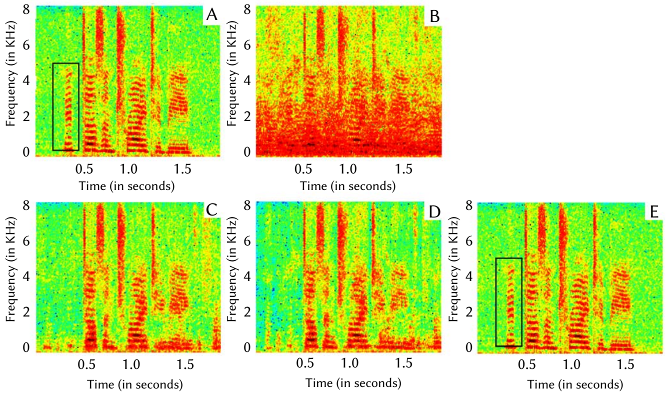


Fig. 6. Spectrotemporal characteristics of the speech processed by different networks.

TABLE VII. Statistical Analysis of Average Intelligibility at 95% Confidence Interval With $F_{Critical}$ Is 3.09 and $P_{Critical}$ Is 0.05

ANOVA SE Models	STOI					
	-5dB		-2dB		0dB	
	P_{Value}	F_{Value}	P_{Value}	F_{Value}	P_{Value}	F_{Value}
CGCRN ⟶ Noisy	<0.001	39.5	<0.001	21.4	<0.001	19.1
CGCRN ⟶ LSTM	<0.001	22.2	<0.002	31.2	<0.001	26.1
CGCRN ⟶ GRU	<0.002	29.4	<0.005	28.6	<0.001	25.3
CGCRN ⟶ CRN	<0.001	21.1	<0.001	19.1	<0.002	18.3

TABLE VIII. Statistical Analysis of Average Quality at 95% Confidence Interval With $F_{Critica}$ Is 3.09 and $P_{Critical}$ Is 0.05

ANOVA SE Models	PESQ					
	-5dB		-2dB		0dB	
	P_{Value}	F_{Value}	P_{Value}	F_{Value}	P_{Value}	F_{Value}
CGCRN ⟶ Noisy	<0.001	32.2	<0.001	37.4	<0.001	22.2
CGCRN ⟶ LSTM	<0.015	18.3	<0.020	28.3	<0.020	27.3
CGCRN ⟶ GRU	<0.001	30.2	<0.001	24.5	<0.001	23.2
CGCRN ⟶ CRN	<0.001	24.1	<0.001	22.1	<0.001	20.5

VI. Summary and Conclusions

In this paper we propose resource efficient Convolutional recurrent network to learn the complex ratio mask for real-time speech enhancement. Convolutional encode-decoder and gated recurrent unit are integrated into the Convolutional recurrent network architecture thereby formulated a causal system, which is suitable for the real-time speech processing. Parallel GRU grouping and efficient skipped connections techniques are used to achieve compact network. Different noise types and speakers are used in training and testing to observe the speaker and noise generalization. With LibriSpeech dataset, the experiments showed that the proposed real-time approach led to improve perceptual speech quality and intelligibility with much fewer trainable parameters than existing LSTM and GRU models. The quality and intelligibility are improved by 31.61% and 17.18% over noisy speech. CGCRN proves comparable spectrotemporal patterns to the clean speech and less distortion can also be noticed. We showed gains on the speech quality and intelligibility with less computational complexity by more effective skip connections and a parallel GRUs structure. The proposed model used fewer parameters and causal operations; therefore, suitable for real-time speech enhancement. The ANOVA statistical analysis confirmed that the intelligibility and quality results are statistically significant. The average STOI test scores are improved from 66.93% to 84.83% with CGCRN (ΔSTOI = 17.90%) in babble noise type. Here Δ indicates the improvements in metrics. Also, the average STOI test scores are improved from 65.60% to 82.05% with CGCRN (ΔSTOI = 16.45%) in cafeteria noise type. On average, 3.09% STOI gain is achieved when the CGCRN when compared to the LSTM network. Moreover, 0.79% improvement in STOI test scores is achieved against the CRN. In addition, the average PESQ test scores are improved from 1.76 to 2.56 with the proposed CGCRN (ΔPESQ = 0.80 equivalent to 31.25%) in babble noise type. at adverse noise levels (-5dB), the CGCRN against the noisy speech (CGCRN ⟶ Noisy), we achieved [F (2, 100) = 39.5, p < 0.001] for STOI and [F (2, 100 = 32.2), p < 0.001] for PESQ, respectively. Also, against the CRN (CGCRN ⟶ CRN), we achieved [F (2, 100) = 21.1, p < 0.001] for STOI and [F (2, 100) = 24.1, p < 0.001] for PESQ. Moreover, against LSTM (CGCRN ⟶ LSTM), we achieved [F (2, 100) = 22.2, p < 0.001] for STOI and [F (2, 100) = 18.3, p < 0.015] for PESQ.

Speech perception quality also depends on the phase. However, since phase lacks spectrotemporal structure, it seems to be impossible to correctly estimate phase spectra using masking-based supervised learning, like in this proposed model. The complex spectral mapping, which concurrently improves the magnitude and phase responses of noisy speech, tries to estimate the real and imaginary spectrograms of clean speech from those of noisy speech. In future work, we would be devoted to proposing more flexible, scalable, and phase included CRNs for real-time speech enhancement, trained on large datasets and tested on the real recordings.

References

[1] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018, doi: 10.1109/TASLP.2018.2842159.

[2] J. Agnew and J. M. Thornton, "Just noticeable and objectionable group delays in digital hearing aids," *Journal of the American Academy of Audiology*, vol. 11, no. 6, pp. 330–336, 2000.

[3] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1381–1390, 2013, doi: 10.1109/TASL.2013.2250961.

[4] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014, doi: 10.1109/TASLP.2014.2352935.

[5] N. Saleem and M. I. Khattak, "Deep Neural Networks for Speech Enhancement in Complex-Noisy Environments," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 1, pp. 84–91, 2020, doi: 10.9781/ijimai.2019.06.001.

[6] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, 2015, doi: 10.1109/TASLP.2015.2512042.

[7] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2014, doi: 10.1109/TASLP.2014.2364452.

[8] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2013, doi: 10.1109/LSP.2013.2291240.

[9] N. Saleem and M. I. Khattak, "Multi-scale decomposition based supervised single channel deep speech enhancement," *Applied Soft Computing*, vol. 95, p. 106666, 2020, doi: 10.1016/j.asoc.2020.106666.

[10] J. Chen and D. Wang, "Long short-term memory for speaker generalization in supervised speech separation," *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. 4705–4714, 2017, doi: 10.1121/1.4986931.

[11] M. Kolbæk, Z. H. Tan, and J. Jensen, "Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 153–167, 2016, doi: 10.1109/TASLP.2016.2628641.

[12] Y. Li and D. Wang, "On the optimality of ideal binary time–frequency masks," *Speech Communication*, vol. 51, no. 3, pp. 230–239, 2009, doi: 10.1109/ICASSP.2008.4518406.

[13] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, Apr. 2015, pp. 708–712, doi: 10.1109/ICASSP.2015.7178061.

[14] K. Tan and D. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 380–390, 2020, doi: 10.1109/TASLP.2019.2955276.

[15] J. Chen, Y. Wang, S. E. Yoho, D. Wang, and E. W. Healy, "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," *The Journal of the Acoustical Society of America*, vol. 139, no. 5, pp. 2604–2612, 2016, doi: 10.1121/1.4948445.

[16] N. Saleem, M. I. Khattak, M. A. Al-Hasan, and A. Jan, "Multiobjective long-short term memory recurrent neural networks for speech enhancement," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 10, pp. 9037–9052, 2021, doi: 10.1007/s12652-020-02598-4.

[17] K. Tan, J. Chen, and D. Wang, "Gated residual networks with dilated convolutions for supervised speech separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Canada, Apr. 2018, pp. 21–25, doi: 10.1109/ICASSP.2018.8461819.

[18] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement," in *Interspeech 2018*, Hyderabad, India, Sep. 2018, pp. 3229–3233, doi: 10.21437/Interspeech.2018-1405.

[19] Z. Zhang, Z. Sun, J. Liu, J. Chen, Z. Huo, and X. Zhang, "Deep recurrent convolutional neural network: Improving performance for speech recognition," *arXiv preprint*, arXiv:1611.07174, 2016.

[20] G. Naithani, T. Barker, G. Parascandolo, L. Bramsl, N. H. Pontoppidan, and T. Virtanen, "Low latency sound source separation using convolutional recurrent neural networks," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2017, pp. 71–75, doi: 10.1109/WASPAA.2017.8169997.

[21] S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," *arXiv preprint*, arXiv:1609.07132, 2016.

[22] S. Jha, A. Dey, R. Kumar, and V. Kumar-Solanki, "A novel approach on visual question answering by parameter prediction using faster region-based convolutional neural network," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 5, pp. 30–38, 2019, doi: 10.9781/ijimai.2018.08.004.

[23] N. Saleem, J. Gao, M. I. Khattak, H. T. Rauf, S. Kadry, and M. Shafi, "Deepresgru: Residual gated recurrent neural network-augmented Kalman filtering for speech enhancement and recognition," *Knowledge-Based Systems*, vol. 238, p. 107914, 2022, doi: 10.1016/j.knosys.2021.107914.

[24] A. Li, C. Zheng, L. Zhang, and X. Li, "Glance and gaze: A collaborative learning framework for single-channel speech enhancement," *Applied Acoustics*, vol. 187, p. 108499, 2022, doi: 10.1016/j.apacoust.2022.108499.

[25] M. I. Khattak, N. Saleem, A. Nawaz, A. A. Almani, F. Umer, and E. Verdú, "ERBM-SE: Extended restricted Boltzmann machine for multi-objective single-channel speech enhancement," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 4, pp. 7–15, 2022, doi: 10.9781/ijimai.2022.03.002.

[26] M. I. Khattak, N. Saleem, J. Gao, E. Verdú, and J. P. Fuente, "Regularized sparse features for noisy speech enhancement using deep neural networks," *Computers and Electrical Engineering*, vol. 100, p. 107887, 2022, doi: 10.1016/j.compeleceng.2022.107887.

[27] T. Gao, J. Du, L. R. Dai, and C. H. Lee, "SNR-based progressive learning of deep neural network for speech enhancement," in *Interspeech 2016*, San Francisco, CA, USA, Sep. 2016, pp. 3713–3717.

[28] A. Laishram and K. Thongam, "Automatic classification of oral pathologies using orthopantomogram radiography images based on convolutional neural network," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 4, pp. 69–77, 2022, doi: https://doi.org/10.9781/ijimai.2021.10.009.

[29] T. Kounovsky and J. Malek, "Single channel speech enhancement using convolutional neural network," in *2017 IEEE International Workshop of Electronics, Control, Measurement, Signals and their Application to Mechatronics (ECMSM)*, Donostia-San Sebastián, Spain, May 2017, pp. 1-5, doi: 10.1109/ECMSM.2017.7945915.

[30] E. M. Grais and M. D. Plumbley, "Single channel audio source separation using convolutional denoising autoencoders," in *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Montreal, QC, Canada, Nov. 2017, pp. 1265–1269.

[31] H. Zhao, S. Zarar, I. Tashev, and C. H. Lee, "Convolutional-recurrent neural networks for speech enhancement," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Canada, Apr. 2018, pp. 2401–2405.

[32] T. Gao, J. Du, L. R. Dai, and C. H. Lee, "Densely connected progressive learning for LSTM-based speech enhancement," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Canada, Apr. 2018, pp. 5054–5058.

[33] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement," in *Interspeech 2018*, Hyderabad, India, Vol. 2018, pp. 3229–3233.

[34] S. Pirhosseinloo and J. S. Brumberg, "Dilated convolutional recurrent neural network for monaural speech enhancement," in *2019 53rd Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, USA, Vol. 2019, pp. 158–162.

[35] N. Manju, B. S. Harish, and N. Nagadarshan, "Multilayer feedforward neural network for Internet traffic classification," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 1, pp. 117–123, 2020.

[36] A. K. Dubey and V. Jain, "Comparative study of convolution neural network's relu and leaky-relu activation functions," in *Applications of Computing, Automation and Wireless Systems in Electrical Engineering: Proceedings of MARC 2018*, Singapore: Springer Singapore, 2019, pp. 873–880.

[37] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *arXiv preprint*, arXiv:1406.1078, 2014.

[38] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, Apr. 2015, pp. 5206–5210.

[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint*, arXiv:1412.6980, 2014.

[40] A. A. Alvarez and F. Gómez, "Motivic pattern classification of music audio signals combining residual and LSTM networks," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 6, pp. 208–214, 2021.

[41] J. G. Beerends, A. P. Hekstra, A. W. Rix, and M. P. Hollier, "Perceptual evaluation of speech quality (PESQ): The new ITU standard for end-to-end speech quality assessment part II: Psychoacoustic model," *Journal of the Audio Engineering Society*, vol. 50, no. 10, pp. 765–778, 2002.

[42] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Dallas, TX, USA, Mar. 2010, pp. 4214–4217.

[43] M. Hasannezhad, Z. Ouyang, W. P. Zhu, and B. Champagne, "An integrated CNN-GRU framework for complex ratio mask estimation in speech enhancement," in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Auckland, New Zealand, Dec. 2020, pp. 642–647.

[44] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," *arXiv preprint*, arXiv:1703.09452, 2017.

Fazal-e-Wahab

Fazal-e-Wahab received a B.E degree in Electronics engineering from the Dawood University of Engineering and Technology, Karachi, Pakistan, in 2009, and the M.Sc. degree in electrical engineering from CECOS University, Peshawar, in 2015. Since 2012, he has been involved in teaching and research with Department of Electrical Engineering, UET Peshawar. Currently, he is pursuing a Ph.D. degree from the University of Science and Technology of China, Hefei, China in Signal and Information Processing. His current research interests are Speech Enhancement, Speech Denoising and Machine learning applications.

Zhongfu Ye

Zhongfu Ye received the BE and MS degrees from the Hefei University of Technology, Hefei, China, in 1982 and 1986, respectively, and the Ph.D. degree from the University of Science and Technology of China, Hefei, China, in 1995. He is currently a Professor of the University of Science and Technology of China. His current research interests are in statistical and array signal processing and image processing.

Nasir Saleem

Nasir Saleem received B.Sc. Telecom Engineering degree from University of Engineering and Technology, Peshawar, Pakistan in 2008, M.S. Electrical Engineering degree from CECOS University, Peshawar, Pakistan in 2012; and Ph.D. degree in Electrical Engineering, specialization in Digital speech processing and Deep Learning from University of Engineering and Technology, Peshawar, Pakistan in 2021. From 2008 to 2012, he was a lecturer at Institute of Engineering Technology, Gomal University, where he was involved in teaching and research. He is now an Assistant Professor in the Department of Electrical Engineering, Faculty of Engineering and Technology, Gomal University. Human-Machine Interaction, Speech Enhancement, and Machine Learning Applications are the areas he is currently researching.

Hamza Ali

Hamza Ali received B.Sc. degree in Electrical engineering from the University of Engineering and Technology, Peshawar, Pakistan, in 2021. He is currently working toward the M.Sc. degree with University of Engineering and Technology, Mardan. His current research interests are broad areas of mobile networking, signal processing, Speech Enhancement and Machine learning.

Imad Ali

Imad Ali received B.Sc. Telecom Engineering degree from the University of Engineering and Technology, Peshawar, Pakistan in 2008, M.S. Electrical Engineering degree from CECOS University, Peshawar, Pakistan in 2012; and a Ph.D. degree in Social Networks and Human-Centered Computing, from National Tsing Hua University, Taiwan, in collaboration with Academia Sinica, Taiwan, in 2020. From 2009 to 2014, he served as a lecturer at different universities in Pakistan. He is currently working as an Assistant Professor in the Department of Computer Science, University of Swat, Pakistan. His research interest includes Question Answering Systems, Data Science, and Machine Learning.

Reading Modi Lipi: A Deep Learning Journey in Character Recognition

Kanchan Varpe, Sachin Sakhare *

Department of Computer Engineering, Vishwakarma Institute of Information Technology, Pune 411048 (India)

* Corresponding author: Kanchanv2007@gmail.com (K. Varpe), sachin.sakhare@viit.ac.in (S. Sakhare)

Received 23 July 2023 | Accepted 21 April 2024 | Published 9 September 2024

Abstract

Advancements in deep learning methodologies have played a significant role in the success of various character recognition processes. Character recognition refers to the technique of identifying either handwritten or printed characters from documents and their conversion into a form that can be read by machines. MODI script, an ancient Indian script, is categorized under the Devanagari script and holds historical significance. Despite its historical importance, there are only a few MODI translators available. Conversely, there exist a vast number of historical documents written in MODI that are yet to be deciphered. Recognizing characters in Indian language scripts poses many challenges due to the complex nature of the scripts and variations in individuals' writing styles. This paper provides an overview of the newest advancements in the Handwritten Optical Character Recognition (HWCR) methodology specifically designed for the MODI script. Utilization of residual networks and inception in image classification has gained popularity in recent times.

In this paper the authors have implemented three techniques: ResNet9, ResNet50, and InceptionNet V3, trained specifically for handwritten MODI characters and vowels. The dataset used for training the models consists of handwritten MODI script images. The benchmark database from IEEE data port for handwritten MODI script is used to evaluate the performance. The dataset contains 46 classes, including 10 vowel classes and 36 consonant classes. Each class comprises 90 images, resulting in a total of 4140 images. The image size in the dataset is 227×227. The accuracy achieved by the trained models is as follows: 98.92% for ResNet9, 91.91% for ResNet50, and 86% for Inception Net V3. The obtained results have been compared with existing models and it is observed that the proposed model attained improved performance parameters and less training and validation losses in comparison to existing methods. There are several advantages of the proposed model in comparison to state of the art, namely minimal training and validation loss. In addition to this, the proposed approach improved generalization and robustness, and improved model scalability.

Keywords

Classification, Deep Learning, Feature Extraction, Image Processing, InceptionNet V3, ResNet50, ResNet9.

DOI: 10.9781/ijimai.2024.09.002

I. Introduction

The script known as "MODI Lipi" is an old script and has significant importance in medieval Maharashtra. It served as the script for a wide range of historical documents, covering subjects such as land revenue, judiciary, justice, religious matters, property matters, and military orders and strategies during the reign of Chhatrapati Shivaji Maharaj and the Peshwas.

For nearly 700 years, MODI remained the medium for administrative and official documentation in Marathi within Maharashtra, until the late 19th century. This prolonged usage emphasizes the need to prioritize the recognition process for the MODI script, as it contains a wealth of valuable information yet to be uncovered.

Researchers are interested in studying handwritten optical character recognition (HOCR) for MODI Lipi, seeking to understand the writing techniques used during the medieval period of the Maratha Empire. The automation of recognition holds the potential to unveil a trove of historical knowledge and understanding. HOCR holds the most promise for the future in image processing, pattern recognition, natural language processing, and analysis of documents. Due to advancements in various technologies for image processing and pattern recognition, considerable improvements were observed in identifying handwritten characters [1].

HOCR is challenging because different scripts have their own style of writing when it comes to shape and continuity. Table I enumerates challenges faced by handwritten Modi character recognition.

Please cite this article as:
K. Varpe, S. Sakhare. Reading Modi Lipi: A Deep Learning Journey in Character Recognition, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 1, pp. 75-83, 2024, http://dx.doi.org/10.9781/ijimai.2024.09.002

A. History

The Modi script has been utilized for writing Marathi language for an estimated duration of 700 years.

The origins of the Modi script are surrounded by various narratives. One theory suggests that Hemadpant, a minister during the rule of Mahadev, is the creator of the Modi script. However, alternative theories propose that Hemadpant did not invent the script from scratch, but rather refined an existing version of Modi before introducing it as the official script for writing Marathi.

According to research conducted by Kulkarni et al. [2], one theory asserts that Hemadpant brought the Modi script from Sri Lanka to India. Another theory suggests that the script was developed during the reign of Chhatrapati Shivaji Maharaj by Baḷaji Avaji, the secretary of state.

These differing accounts contribute to the intriguing history and origins of the Modi script, highlighting the need for further research and exploration to uncover its true genesis and evolution.

TABLE I. Challenges for Handwritten MODI Character Recognition

Sr. No.	Challenges
1	Large number of classes.
2	Open and closed loops, arcs, strokes, straight lines.
3	Various strokes existing in a character may touch each other due to hasty writing.
4	Large intra-class variations due to different writing styles.
5	There are inter-class similarities.
6	Extracting structural features is very difficult due to the complex structure of some of its alphabet letters.
7	MODI is written in cursive type; hence it creates extra branches in letters.

Modi evolved into different forms over the years until the 20th century when the Balbodh style of Devanagari script was adopted as a standard form for writing Marathi. Different styles of Modi are listed according to the eras and centuries in which they emerged in Table II. As all documents before the 20th century were written in the Modi script, this makes Modi historically significant [3].

TABLE II. Modi Script Writing Forms Over Years

Modi style	Time Period
Proto-Modi (आद्यकालीन)	12th Century
Yadav Era (यादवकालीन)	13th Century
Bahamani Era (बहमनीकालीन)	14th – 16th Century
Shiva Era (शविकालीन)	17th Century
Chitnisi (चतिनीसी)	17th Century
Peshwa Era (पेशवेकालीन)	Lasted till 1818
Anglakalin or British Colonial Era (आंग्लकालीन)	1818-1952

B. Properties of MODI Script

During the 12th century, the MODI script gained prominence as a writing system for the Marathi language. This ancient script continued to be extensively utilized starting in the 12th century. The term "MODI" is deemed to have been derived from the Marathi verb "moḍaṇe," which means to "to break or bend" [4].

To write in the MODI script, a writing instrument called "Boru" or "Lekhan" was utilized, with the pen being made from bamboo. Despite the fact that the MODI script is based on the Devanagari script, there are significant variations between them. The MODI script is part of the Nagari family of scripts and is most suitable for continuous writing. These variations show up in rendering behaviors, letter forms and orthography of the characters. Joseph et al. [5] provide a comprehensive exploration of these distinguishing features.

The unique characteristics and historical significance of the MODI script have attracted considerable scholarly attention, prompting in-depth investigations into its structural intricacies and functional aspects.

The behaviors exhibited by characters in specific contexts, such as combinations of consonants and vowels and consonant conjuncts, are intrinsic features of MODI orthography and differentiate it from the Devanagari script. The MODI script comprises 46 distinct letters, with 10 vowels and 36 consonants. The task of word segmentation for the MODI script is quite challenging as no termination symbol is used to designate the end of a sentence. In the MODI script, all the characters are written as if they are hanging from a horizontal line that is drawn across the page. This distinct style sets MODI apart. Notably, research conducted by Kulkarni et al. [6] highlights that the elimination of terminating symbols in the MODI script significantly increased the writing speed.

This removal reduced the need to lift the "Boru" pen frequently. The unique characteristics of the MODI script play a crucial role in defining its identity and require specialized approaches for precise interpretation and analysis. Continual research endeavors to further enhance our comprehension of MODI orthography and foster the development of effective methods for working with MODI script documents.

By delving into the intricacies of the MODI script, researchers strive to uncover its nuances, intricacies, and underlying patterns. This deeper understanding helps in refining techniques for accurate recognition, transcription, and translation of MODI script content. Additionally, ongoing research contributes to the advancement of tools and methodologies that facilitate efficient handling and processing of MODI script documents.

The ultimate goal is to bridge the gap between the historical significance of the MODI script and its effective utilization in contemporary contexts. By harnessing insights gained from ongoing research, scholars, language enthusiasts, and technological advancements can collaborate to preserve, analyze, and leverage the rich cultural heritage contained within MODI script documents.

Modi is a beautifully flowing cursive script known for its elegance. The fundamental graphical structure of Modi letters, as shown in Fig. 1, Fig. 2, and Fig. 3, follows the traditional bārākhadi format.

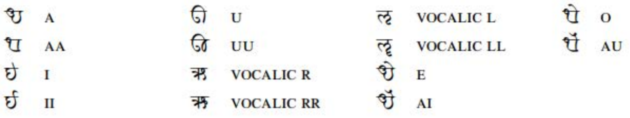

A
AA
I
II
U
UU
VOCALIC R
VOCALIC RR
VOCALIC L
VOCALIC LL
E
AI
O
AU

Fig. 1. Vowels used in Modi Script.

There are 13 combining vowel signs as shown in Fig. 2.

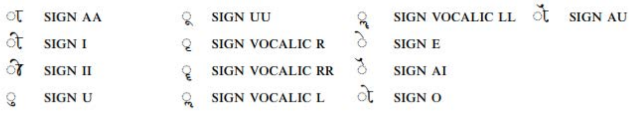

SIGN AA
SIGN I
SIGN II
SIGN U
SIGN UU
SIGN VOCALIC R
SIGN VOCALIC RR
SIGN VOCALIC L
SIGN VOCALIC LL
SIGN E
SIGN AI
SIGN O
SIGN AU

Fig. 2. Vowel Signs in Modi Script.

There are 34 consonant letters as shown in Fig. 3.

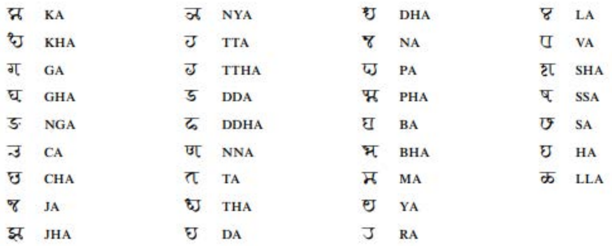

KA
KHA
GA
GHA
NGA
CA
CHA
JA
JHA
NYA
TTA
TTHA
DDA
DDHA
NNA
TA
THA
DA
DHA
NA
PA
PHA
BA
BHA
MA
YA
RA
LA
VA
SHA
SSA
SA
HA
LLA

Fig. 3. Consonants in Modi Script.

The organization of the remainder of the paper is set out here. Section II contains an in-depth and organized summary of a literature review of the innovative methodologies that researchers have proposed along with challenges from state of the art. Section III contains an outline of research techniques of the proposed Convolutional Neural Network (CNN)-based deep learning models. Section IV describes the database. Section V includes a discussion of the initial outcome procured from the experimentation utilizing the deep learning architectures. Finally, Section VI concludes and highlights future research possibilities.

II. Literature Review

Table III presents various research works that focus on MODI character or numeral recognition. Each researcher has employed distinct methodologies in their studies.

In the research conducted by Besekar et al. [7], classification and recognition of MODI characters were achieved using mathematical morphology and decision tree algorithms. Features extracted from vertical and horizontal lines, blobs and concavities were utilized, resulting in an accuracy of 75%.

Besekar et al. [8] employed a two-layer feed-forward neural network with scaled conjugate gradient for MODI vowel classification. Different accuracies were obtained by using different directional chain codes. For instance, 65.3% accuracy was achieved using the 4-directional chain code histogram (CCH4) and the 4-directional normalized chain code histogram (CH4D). Furthermore, using the 4-directional CCH, the 4-directional NCCH, and a centroid, the accuracy improved to 67.9%. Similarly, the 8-directional CCH and the 8-directional NCCH resulted in 65.9% accuracy, while utilizing the 8-directional CCH, the 8-directional NCCH, and a centroid increased the accuracy to 73.5%.

Ramteke et al. [9] introduced the use of a Variance table to categorize MODI numerals, achieving an accuracy of 93.5%. This approach involved dividing the numeral into four identical 15 × 15 square zones and analyzing their polar coordinates, variance, theta, and rh distance to generate a feature set for recognition.

These research studies demonstrate various techniques and algorithms employed for MODI character recognition, highlighting the progress and achievements in this field.

In the study conducted by Katkar et al. [10], the utilization of the Kohonen Neural Network for classification, coupled with a measured structural similarity approach for feature extraction, resulted in a performance rate ranging from 91% to 97%.

The Euclidean distance classifier is generally employed for MODI character identification. In the research by Kulkarni et al. [5], a correct recognition rate of 94.92% was obtained using Zernike moments, while an accuracy of 94.78% was obtained using Zernike complex moments with an integrated approach for heterogeneous zones in offline character recognition.

Kulkarni et al. [11] employed Hu's seven invariant moments as the feature vector for training samples, and recognition of test sample features was performed using the Euclidean distance classifier. The accuracy achieved with Hu's invariant seven moments was 70%. Zernike Moments, on the other hand, provide statistical measures of pixel distribution around the character's center of gravity and capture information at a single boundary point. When using Zernike moments, an accuracy of 86.66% was obtained.

In the research work of Sidra Anam et al. [12], the Kohonen Neural Network was employed for classification, preceded by Otsu's Binarization algorithm, resulting in a performance rate of 72.6%. Otsu's threshold algorithm was employed for the Modi Script Character Recognizer System (MSCR).

Deshmukh et al. [13], utilized the chain code feature extraction technique with a non-overlapping blocking strategy, followed by the use of correlation coefficient. The correlation feature (r or R) is a measure of the strength and direction of the linear relationship between two variables. It is calculated by dividing the covariance of the variables by the product of their standard deviations. The maximum recognition rate procured on a database of 30,000 images was 85.21%. The results of the recognition showed an improved performance when using 5 × 5 grid divisions. These research studies highlight various approaches and techniques employed for MODI character recognition, showcasing the advancements and achievements in this field.

Chandure et al. [14], conducted a study in which the entire character set of the MODI script was recognized using the chain code in combination with BPNN, KNN, and SVM. The respective accuracies obtained were 37.5%, 60%, and 65%. Feature extraction was performed using these techniques while combining the BPNN, KNN, and SVM techniques, resulting in accuracies of 15%, 40%, and 47.5% respectively. There are multifarious reasons of low accuracy namely inadequate training of model such as imbalanced dataset and limited discriminative power of algorithm. etc.

The output layer consisted of 13 neurons for Devanagari characters and 12 neurons for MODI characters. When the chain code feature extraction technique was used, an accuracy of 37.5% was achieved, while an accuracy of 15% was obtained when the insertion junction feature extraction technique was employed.

Since there are 13 classes of vowels, 13 SVM classifiers were required for their separation. When the chain code was used as a feature extraction technique, an accuracy of 65% was achieved. Similarly, when the insertion junction was used as a feature extraction technique, an accuracy of 47.5% was obtained.

The k-nearest neighbor algorithm (KNN) yielded an accuracy of 60% when the chain code feature extraction technique was employed, and an accuracy of 40% when the insertion junction feature extraction technique was utilized.

In research conducted by Gharde et al. [15], a combination of moment invariant and affine moment invariant approaches was used. This hybrid approach extracted 18 features from each number or character. SVM was employed as the classifier, resulting in an accuracy of 89.72%. An accuracy of 89.72% was obtained when two feature extraction approaches, namely, moment invariant and affine moment invariant were utilized along with SVM being used as the classifier.

Maurya et al. [16] came up with a proposal for a framework for recognizing handwritten MODI characters digitally. The authors used heuristics that were empirically determined to figure out the contribution of features from a hybrid feature space for character recognition. The pre-processing methods included noise removal, binarization, skeletonization, character segmentation and smoothing. The characters that were segmented underwent post-processing and recognition once the pre-processing was done. The average accuracy obtained was 91.20%, with a claimed best accuracy of 99.10%.

TABLE III. Overview of Research Work Related to Modi Character Recognition

Sr. No.	Author	Feature Extraction	Classification	Data type			Accuracy%
				Characters	Vowels	Numerals	
(1)	Besekar D.N. et al., 2011 [7]	Blobs, vertical & horizontal lines, concavities	Mathematical morphology, decision tree.	-	-	✓	75
(2)	Besekar D.N. et al., 2012 [8]	Chain code, image, and centroid method	Two-layer feed forward network with scaled conjugate gradient.	-	✓	-	65.3, 67.9, 65.9, 73.5
(3)	Ramteke R.J. et al., 2012 [9]	Polar coordinate of zone, Variance, theta angle and Rh distance	Comparing the variance table.	-	-	✓	93.5
(4)	Katkar G. S. et al., 2013 [10]	Structural similarity	Kohonen neural network, BPNN	✓	-	-	91-97
(5)	Solley Joseph et al., 2014 [5]	Zoning, (1) Zernike Moments and (2) Zernike complex moments	Euclidean Distance	✓	-	-	1) 94.92 2) 94.78
(6)	Kulkarni S.A et al., 2015 [11]	(1) Hu's invariant features (2) Zernike moments	Euclidean distance	-	-	✓	1) 70 2) 86.66
(7)	Sidra Anam et al., 2015 [12]	-	Kohonen neural network	✓	-	-	72.6
(8)	Manisha S. et al., 2015 [13]	Chain code feature extraction and non-overlapping blocking strategy	Correlation coefficient	-	-	✓	85.21
(9)	Chandure S.L. et al., 2016 [14]	Chain Code Histogram Features, Intersection / Junc. Features	1)BPNN 2)KNN 3)SVM	✓	-	-	Chain Code: 60, 37.5, 65 InsertionJunc 40, 15, 47.5
(10)	Gharde S.S. et al., 2016 [15]	Using Moment Invariant and Affine Moment Invariant	SVM	✓	-	-	89.72%
(11)	Maurya R.K. et al., 2018 [16]	Chain code	Empirically determined heuristics	✓	-	-	Average: 91.20% Best: 99.10%
(12)	Joseph S. et al., 2020 [17]	-	1)Euclidean distance classifier, 2)Manhattan distance classifier	✓	-	-	1) 99.28% 2) 94%
(13)	S. Joseph et al., 2020 [18]	CNN autoencoder	SVM	✓	-	-	99.3%
(14)	Shruti Sawant et al., 2020 [19]	-	CNN	✓	-	-	95.44%, 95.97%
(15)	Solly Joesph et al., 2021[20]	WT-SVD	Euclidean distance	✓	-	-	99.5%
(16)	Joseph S. et al., 2021 [21]	-	ACNN	✓	-	-	99.78%
(17)	Tamhankar et al., 2021 [22]	-	DCNN	✓	-	-	64%
(18)	Kirti et al., 2021 [23]	-	AlexNet	✓	-	-	89.72 %
(19)	Chandure et al., 2021 [24]	-	1)Supervised TL 2)SVM	✓	-	-	92.32%
(20)	Kulkarni S.A. et al., 2021 [25]	Zoning, 1.Zernike moments 2.Zernike Complex moments 3.Ensemble Bagging	Euclidean distance classifier	✓	-	-	1. 94.92% 2. 94.78% 3. 97.68%
(21)	Jidnyasa Kondhare et al., 2022 [26]	-	1)CNN 2)VGG16	✓	-	✓	1)76.46% 2)92.48%
(22)	Maitreyi Ekbote et al., 2022 [27]	-	1)Random Forest 2)XGBoost	✓	-	✓	1)92% 2)93.3%
(23)	Vishal Pawar et al., 2022 [28]	-	CNN	✓	-	-	91.62%
(24)	Chaitali Chandankhede et al., 2023 [29]	-	1)Inception V3 2)ResNet 50	✓	✓	-	93.923% 94.552%

In the study conducted by Joseph et al. [17], two algorithms utilizing distance classifiers were implemented for the classification of handwritten Modi script. The data underwent vectorization, followed by noise reduction techniques. The first experiment utilized Euclidean distance classifiers, while the second experiment used the Manhattan distance classifier. The procured accuracies were 99.28% and 94% respectively. The Manhattan distance method demonstrated better performance in terms of time complexity, although it was less accurate in comparison to the second method.

S. Joseph et al. [18], came up with a proposal to make use of a CNN autoencoder as a feature extractor in order to recognize characters in the MODI script. Making use of the CNN autoencoder, the feature set size was decreased from 3600 to 300. SVM was then used as a classifier for the features that were extracted and an accuracy of 99.3% was achieved.

Two architectures were considered in the study conducted by Shruti Sawant et al. [19]. The first framework was made up of two convolution layers that came first. Next came max pooling and fully connected layers. The second architecture comprised three sets of convolution and max pooling layers. Three fully connected layers followed. The accuracy obtained for the first architecture was 95.44%, and for the second architecture, it was 95.97%.

Joseph et al. [21], utilized an Augmented CNN (ACNN) model by incorporating data augmentation techniques such as 45-degree horizontal and vertical flips. This combined approach with a CNN resulted in an accuracy of 99.78%.

Tamhankar P.A. et al. [22], implemented a Deep CNN (DCNN) with Rectified Linear Unit neural activation in the convolutional layers, which improved the performance and reduced computational requirements. Their work achieved an accuracy of 64%.

Mahajan Kirti et al. [23], employed a pre-trained Conventional Neural Network called AlexNet for training their model. AlexNet has been trained on a vast dataset of 15 million labeled high-resolution images from 22,000 categories. The experimental arrangement utilizing the AlexNet model in MATLAB yielded an accuracy of 89.72%.

Savitri Chandure et al. [24], utilized a DCNN, namely, AlexNet as a pre-trained network, transferring its weights for retraining. They trained a SVM classifier on activation features to procure classifier models, obtaining an accuracy of 92.32%.

Kulkarni S.A et al. [25] used Zernike moments to achieve an accuracy of 94.92% for an integrated approach, while Zernike Complex moments yielded 94.78% for the integrated approach.

Jidnyasa Kondhare et al. [26] employed CNN for Modi character recognition, achieving a training accuracy of 73.93%, a validation accuracy of 76.46%, and a training time of 9866 seconds. They also tried VGG16, which resulted in an accuracy of 99.73% for training, an accuracy of 92.48% for validation, and a training time of 7267 seconds.

Maitreyi Ekbote et al. [27] proposed a character recognition model on the basis of a CNN for effectively identifying MODI numerals and alphabets. They enhanced the model by utilizing a classifier like Random Forest or XGBoost, achieving recognition accuracies of 92% for characters and 93.3% for numerals.

Vishal Pawar et al. [28] developed a CNN model to recognize characters in the MODI script. Due to the limited dataset of 4140 images, they applied data augmentation methodologies such as flipping, rotation, noising, and blurring to expand the dataset. The trained model was approximately 91.62% accurate in recognizing handwritten MODI characters.

Chaitali Chandankhede et al. [29], experimented with character recognition using ResNet50 and InceptionNet V3. Their dataset consisted of about 8000 photos, and they employed the Global Otsu threshold approach for binarization. The processed images recognized using ResNet50 achieved a testing accuracy of 94.552% with a model precision of 0.86, while InceptionV3 provided a testing accuracy of 93.923% with a model precision of 0.843. The article suggests further research in different binarization strategies, varied CNN models, regularization treatment configurations, and the automation of a powerful word recognition model.

III. Methodology

The primary focus of this work is to leverage deep learning models, specifically Residual Networks and Inception V3, for character recognition in the MODI script. To evaluate performance, the benchmark database for handwritten MODI script from IEEE DataPort [30] is utilized.

The training involves handwritten MODI script images from this IEEE DataPort dataset, which serves as the benchmark for model performance evaluation. This dataset includes 46 distinct classes: 10 vowel classes and 36 consonant classes. Each class contains 90 images, leading to a total of 4140 images. These images are standardized to a size of 227×227 pixels.

The hyperparameters are as follows: an initial learning rate of 0.001, a batch size of 32, activation functions including Sigmoid and ReLU, and the Adam optimizer.

A. ResNet 9-Residual Networks

ResNet is short for Residual Networks, and in our model, we pass our data through 9 layers. In Residual Networks, skip connections are employed for connecting the activations of each of the layers to the other layers, omitting a few layers in between. This forms a residual block, and multiple residual blocks are stacked on top of each other to create the entire network. In our character recognition model, we have 8 convolutional blocks. Each block performs a convolution operation on the image using a 3x3 kernel size, followed by batch normalization. The results of each block are then passed through the ReLU activation function and sent to the next block. The skip connections are utilized to form a residual block, which consists of two consecutive convolutional blocks. Additionally, max pooling is performed to down sample or pool the feature map before passing it to the next layers. The flow of this process is illustrated in Fig. 4.

We will utilize the ResNet-9 architecture in our character recognition model. In our model, we have 8 convolutional blocks, each performing a convolution operation on the image using a 3x3 kernel size, followed by batch normalization. The output of each block is normalized and then passed through the ReLU activation function before being sent to the next block. A residual block is formed by stacking two consecutive convolutional blocks, and skip connections are used to connect the activations.

Additionally, max pooling is performed after each block to down sample or pool the feature map before passing it to the next layers, creating a down sampled or pooled feature map. This flow is illustrated in Fig.4.

To classify images based on predictions, our ResNet model utilizes max pooling to reduce the spatial dimensions, followed by flattening the feature vector into a linear vector. This linear vector is then passed through a dropout layer to consider only relevant features before being fed into the linear layer. The linear layer performs a linear transformation to obtain the class probabilities vector. The predicted class is determined by selecting the class with the highest probability. This process is depicted in the classifier block shown in Fig. 4.

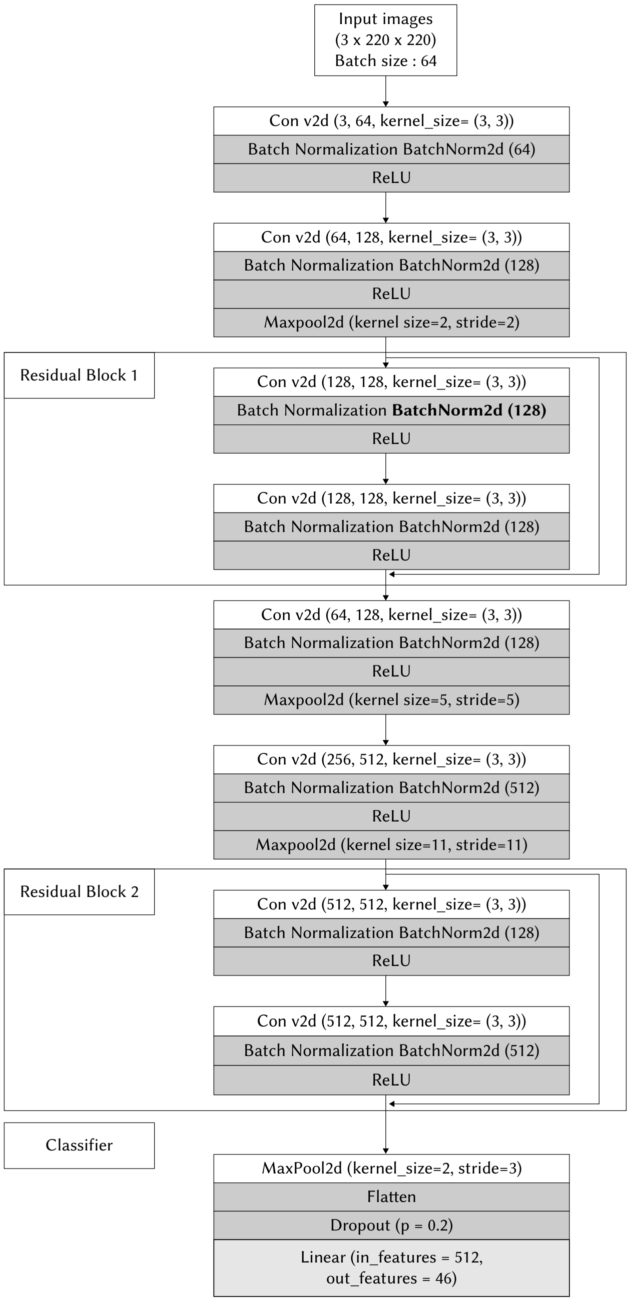

Input images
(3 x 220 x 220)
Batch size : 64
Con v2d (3, 64, kernel_size= (3, 3))
Batch Normalization BatchNorm2d (64)
ReLU
Con v2d (64, 128, kernel_size= (3, 3))
Batch Normalization BatchNorm2d (128)
ReLU
Maxpool2d (kernel size=2, stride=2)
Residual Block 1
Con v2d (128, 128, kernel_size= (3, 3))
Batch Normalization BatchNorm2d (128)
ReLU
Con v2d (128, 128, kernel_size= (3, 3))
Batch Normalization BatchNorm2d (128)
ReLU
Con v2d (64, 128, kernel_size= (3, 3))
Batch Normalization BatchNorm2d (128)
ReLU
Maxpool2d (kernel size=5, stride=5)
Con v2d (256, 512, kernel_size= (3, 3))
Batch Normalization BatchNorm2d (512)
ReLU
Maxpool2d (kernel size=11, stride=11)
Residual Block 2
Con v2d (512, 512, kernel_size= (3, 3))
Batch Normalization BatchNorm2d (128)
ReLU
Con v2d (512, 512, kernel_size= (3, 3))
Batch Normalization BatchNorm2d (512)
ReLU
Classifier
MaxPool2d (kernel_size=2, stride=3)
Flatten
Dropout (p = 0.2)
Linear (in_features = 512,
out_features = 46)

Fig. 4. Resnet 9's architecture.

B. Resnet 50

ResNet-50 [31] is a CNN with 50 layers, and it is a widely used architecture in computer vision applications. ResNet, short for Residual Networks, is a common neural network that has enabled the training of deep networks with hundreds of layers. In the ResNet architecture, denoted as the Residual Neural Network, the labels "50 layers," "101 layers," and "152 layers" correspond to the collective count of convolutional layers integrated into the network. These numerical designations, encapsulated within square brackets like [3, 4, 5, 6], [3, 4, 23, 8], and [3, 8, 36, 3], signify the arrangement of convolutional layers within distinct segments of the ResNet structure.

One of the challenges faced by CNNs is the "Vanishing Gradient Issue," where gradients significantly diminish during back propagation, leading to minimal weight updates. To address this problem, ResNet introduces a solution known as "SKIP CONNECTION."

A skip connection is a direct connection that bypasses certain layers in the model. This skip connection alters the output. When a skip connection is not used, the input 'X' is multiplied by the layer weights, and a bias term is added, followed by the activation function F () to obtain the output as:

$$F\,(w*x + b\,(\text{equivalent to } F(X)) \tag{1}$$

However, with the skip connection technique, the output becomes:

$$F(X) + X \tag{2}$$

Before training the model, preprocessing steps are applied to the images. This involves padding and cropping the images to a size of (224,224) to ensure uniformity.

The images are also adjusted through normalization. The standard deviation of each channel is divided by the means of the image tensors before subtracting them.

This normalization ensures that values from any one channel do not disproportionately influence losses and gradients during training.

To classify images based on predictions, our ResNet model utilizes max pooling to downs ample the feature map, followed by flattening the resulting feature vector into a linear vector. A dropout layer is applied to retain only relevant features before passing them to a linear layer, which performs a linear transformation to obtain the class probabilities vector. The predicted class is determined based on the maximum probability.

C. InceptionNet V3

Inception-v3 is a CNN consisting of 48 layers. It is known for its effective architecture design and parameter reduction techniques. The network incorporates three different kinds of Inception modules: Inception A, Inception B, and Inception C. These modules combine both convolutional layers and pooling layers in a parallel fashion to capture distinctive features while reducing the number of parameters. To achieve parameter reduction, small convolutional layers such as 3×3, 1×3, 3×1, and 1×1 are employed within the Inception modules. The model also includes symmetric and asymmetric components like convolutions, average pooling, max pooling, concatenations, dropouts, and fully connected layers. Batch normalization is utilized to a great extent, ensuring improved training stability. Softmax is employed to compute the loss during training.

InceptionNet V1 serves as the foundation for subsequent versions, including Inception-v3. Each version builds upon the previous one, introducing iterative improvements. In the case of Inception-v3, notable modifications include factorization into smaller convolutions, spatial factorization using asymmetric convolutions, the incorporation of auxiliary classifiers, and efficient grid size reduction. For our specific implementation, we adjusted the number of output channels in the final layer from 1,000 to 98, as our network had fewer classes compared to the original Inception-v3 model. The original model was designed for a dataset with 1,000 classes. InceptionNet performs exceptionally well on large datasets, and expanding our dataset can enhance the model's accuracy in comparison to other models [32].

IV. Database Description

This section of the paper includes a description of datasets that are available publically for the handwritten MODI script [30]. System performance is analyzed on the benchmark database from the IEEE data port (https://ieeedataport.org/documents/handwritten-modi-characters). The dataset for handwritten MODI script comprises 46 classes, with 10 classes for vowels and 36 classes for consonants. Each class contains 90 images, resulting in a total of 4140 images. To

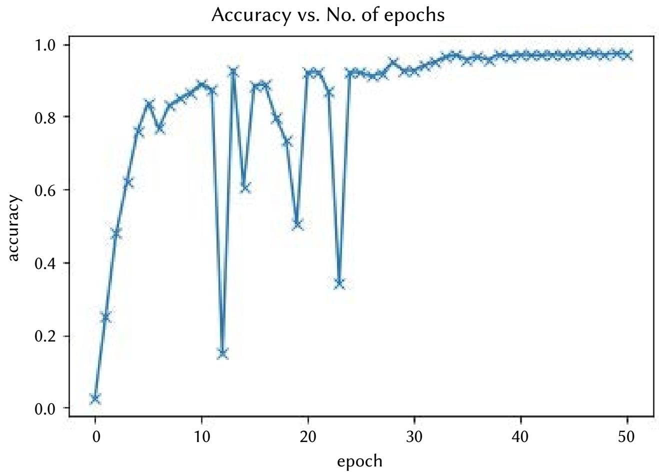


Fig. 5. ResNet 9 Accuracy Plot.

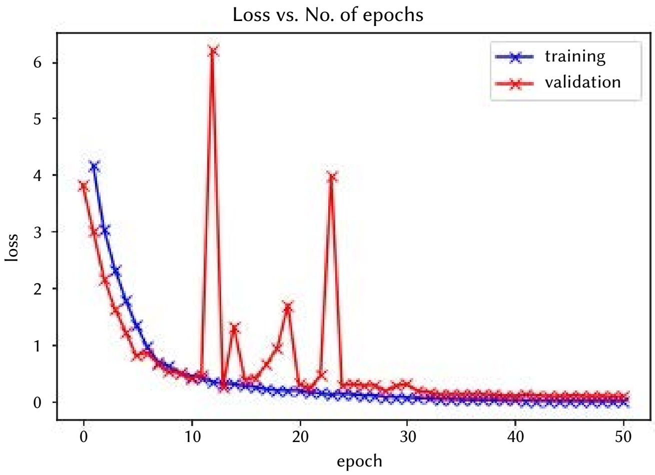


Fig. 6. ResNet 9 Loss Plot.

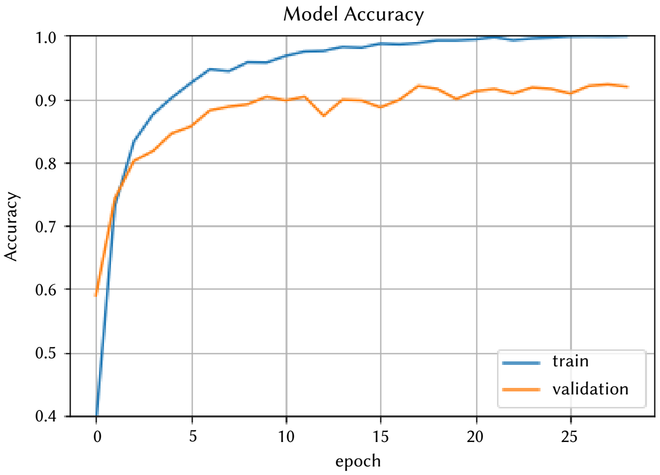


Fig. 7. ResNet 50 Accuracy Plot.

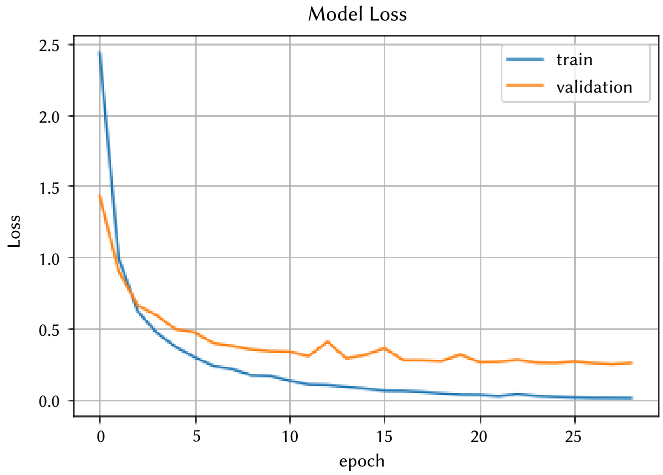


Fig. 8. ResNet 50 Loss Plot.

prepare the dataset for training and testing, it was split into a ratio of 80:20, generating 3336 training images and 834 testing. images. During training, a batch size of 32 was utilized, along with other hyper parameters including weight decay, gradient clipping, and the Adam optimizer as the optimization function with a specified learning rate.

In this research, there was employed multifarious data augmentation methods to improve the diversity as well as robustness of the dataset utilized for model training. These techniques include but not limited to, image rotation, vertical and horizontal flipping, random cropping and zooming, brightness and contrast adjustment. Arbitrary rotation of pictures by a certain degree is also done to simulate alterations in image perspective.

Images were mirrored vertically or horizontally to augment the dataset with additional variations. Arbitrary zooming and cropping were applied to introduce changes in image scaling and composition. Brightness and contrast levels were adjusted to simulate different lighting conditions.

V. Experimental Results and Discussion

All the mentioned CNN models were trained with different values of hyper parameters to get optimal results. The performances of these CNN models are d iscussed in this section.

A. ResNet9

Data augmentation techniques were utilized to enhance the generalization and training effectiveness of the model. The images were converted into tensors using the PyTorch framework for training purposes.

The model underwent training for 50 epochs, resulting in a training accuracy of 97.32%. The accuracy progression throughout the epochs is illustrated in Fig. 5. Additionally, the model's loss consistently decreased over time.

Subsequently, the trained model was evaluated using the test dataset images, yielding a test accuracy of 98.92% shown in Fig. 6.

B. ResNet50

The ResNet-50 model underwent training for 29 epochs to achieve optimal accuracy on the dataset. The training accuracy obtained was 99.94%, while the testing accuracy reached 91.91%.

Fig. 7 illustrates the relationship between the model accuracy and the number of epochs, showcasing how the accuracy improves over time. On the other hand, Fig. 8 showcases the model loss throughout the epochs, demonstrating a gradual decrease in loss.

C. InceptionNet V3

The Inception V3 model underwent training for 50 epochs. The training accuracy achieved was 78.6%, while the testing accuracy reached 86%.

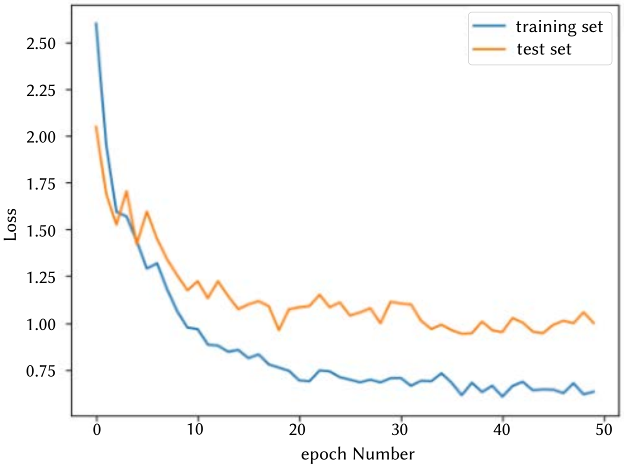


Fig. 9. Inception V3 Loss Plot.

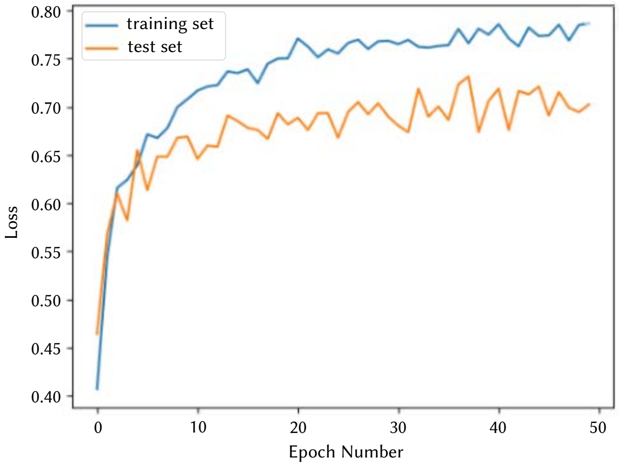


Fig. 10. Inception V3 Accuracy Plot.

Fig. 9 illustrates how training and validation accuracy and the number of epochs are related, providing insights into how accuracy evolves during training. Additionally, Fig. 10 showcases the training and validation loss throughout the epochs, demonstrating the gradual decrease in loss.

TABLE IV. Training and Testing Accuracy

Technique used	Number of Epochs	Training Accuracy	Testing Accuracy
ResNet-9	50	97.32%	98.92%
ResNet-50	29	99.94%	91.91%
InceptionNet V3	50	78.6%	86%

The models demonstrated significant performance levels. ResNet-50, trained over 29 epochs, achieved a training accuracy of 99.94% and a testing accuracy of 91.91%. Inception V3, after its training process, recorded a training accuracy of 78.6% and a testing accuracy of 86%, as presented in Table IV. Meanwhile, in our experiments, ResNet-9 yielded a training accuracy of 97.32% and a testing accuracy of 98.92%

VI. Conclusion and Future Scope

The work in this paper focuses on utilizing deep learning models, specifically Residual Networks and Inception V3, for MODI script character recognition. The training process involved fine-tuning various hyperparameters to obtain the desired outcomes. However, challenges such as a limited image dataset, similarities between classes, word continuity, and the cursive nature of the MODI script remain significant obstacles in handwritten MODI character recognition. To address these challenges, future work can focus on augmenting the dataset to avoid inter-class misclassification and improve CNN invariance qualities. Additionally, there is potential for enhancing recognition accuracy and expanding the scope to include word and line recognition. Segmenting text in the MODI script poses a considerable challenge due to the absence of word or sentence stopping symbols, making it an area for future investigation.

References

[1] V. Kanchan and S. Sakhare, "Review of Character Recognition Techniques for MODI Script," *Indian Journal of Science and Technology*, vol. 16, no. 26, pp. 1935-1946, 2023. doi: 10.17485/IJST/v16i26.485.

[2] K. Sadanand, L. Prashant, R. Ramesh and L. Pravin, "Impact of zoning on Zernike moments for handwritten MODI character recognition," *2015 International Conference on Computer, Communication and Control (IC4), Indore*, India, 2015, pp. 1-6, doi: 10.1109/IC4.2015.7375516.

[3] S. Joseph and J. George, "Handwritten Character Recognition of MODI Script using Convolutional Neural Network Based Feature Extraction Method and Support Vector Machine Classifier," *2020 IEEE 5th International Conference on Signal and Image Processing (ICSIP), Nanjing, China*, 2020, pp. 32-36, doi: 10.1109/ICSIP49896.2020.9339435

[4] A. Pandey, "Proposal to Encode North Indic Number Forms in ISO/IEC 10646," 2007. [Online]. Available: https://api.semanticscholar.org/CorpusID:215863507

[5] S. Joseph and J. George, "Feature Extraction and Classification Techniques of MODI Script Character Recognition," *Pertanika Journal of Science and Technology*, vol. 27, pp. 1649-1669, 2019. Available: https://api.semanticscholar.org/CorpusID:260481222.

[6] K. Sadanand, P. Borde, R. Ramesh, and P. Yannawar, "Offline MODI Character Recognition Using Complex Moments," *Procedia Computer Science*, vol. 58, pp. 516-523, 2015, doi: 10.1016/j.procs.2015.08.067.

[7] D. Besekar, "Recognition of numerals of MODI script using morphological approach," *International Referred Research Journal*, ISSN 0974-2832, pp. 0974-2832, 2011

[8] D. Besekar, "Special Approach for Recognition of Handwritten MODI Script's Vowels," *in National Conference "MEDHA 2012"*, vol. 1, no. 1, pp. 48-52, September 2012. [Online]. Available: /proceedings/medha/number1/8679-1023/.

[9] D. Besekar and R. Ramteke, "Feature extraction algorithm for handwritten numerals recognition of MODI script using zoning-based approach," *International Journal of Systems, Algorithms and Applications,* vol. 2, pp. 1-4, 2012

[10] S. Ramteke and G. Katkar, "Recognition of Off-line Modi Script: A Structure Similarity Approach," *International Journal of Research in Engineering, IT and Social Science (IJREISS)*, vol. 2, no. 11, pp. 2250-0588 2013. [Online]. Available: https://api.semanticscholar.org/CorpusID:16802396

[11] S. Kulkarni and P. Borde, "Analysis of orthogonal moments for recognition of handwritten MODI numerals," *VNSGU Journal of Science and Technology*, vol. 4, pp. 36-43, Jul. 2015.

[12] S. Anam and S. Gupta, "An approach for recognizing Modi Lipi using Ostu's binarization algorithm and Kohenen neural network," *International Journal of Computer Applications*, vol. 111, no. 2, pp. 29–34, Feb. 2015, doi: 10.5120/19511-1128.

[13] M. Deshmukh, M. Patil, and S. Kolhe, "Off-line handwritten Modi numerals recognition using chain code," *in Proceedings of the Third International Symposium on Women in Computing and Informatics, Kochi, India*, 2015, pp. 388–393, doi: 10.1145/2791405.2791419.

[14] S. Chandure and V. Inamdar, "Performance analysis of handwritten Devnagari and MODI Character Recognition system," *2016 International Conference on Computing, Analytics and Security Trends (CAST), Pune, India*, 2016, pp. 513-516, doi: 10.1109/CAST.2016.7915022.

[15] S. Gharde and R. Ramteke, "Recognition of characters in Indian MODI script," *2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC), Jalgaon, India,* 2016, pp. 236-240, doi: 10.1109/ICGTSPICC.2016.7955304

[16] R. Maurya and S. Maurya, "Recognition of a Medieval Indic-Modi Script

using Empirically Determined Heuristics in Hybrid Feature Space," *International Journal of Computer Sciences and Engineering*, vol. 6, no. 2, pp. 136-142, Feb. 2018, doi: 10.26438/ijcse/v6i2.136142.

[17] S. Joseph, J. George, and S. Gaikwad, "Character Recognition of MODI Script Using Distance Classifier Algorithms," *in ICT Analysis and Applications*, S. Fong, N. Dey, and A. Joshi, Eds. Singapore: Springer Singapore, 2020, pp. 105–113.

[18] S. Joseph and J. George, "Handwritten Character Recognition of MODI Script using Convolutional Neural Network Based Feature Extraction Method and Support Vector Machine Classifier," 2020 *IEEE 5th International Conference on Signal and Image Processing (ICSIP), Nanjing, China*, 2020, pp. 32-36, doi: 10.1109/ICSIP49896.2020.9339435

[19] S. Sawant, A. Sharma, G. Suvarna, T. Tanna and S. Kulkarni, "Word Transcription of MODI Script to Devanagari Using Deep Neural Network," *2020 3rd International Conference on Communication System, Computing and IT Applications (CSCITA), Mumbai, India*, 2020, pp. 18-22, doi: 10.1109/CSCITA47329.2020.9137781

[20] J. Solley and J. George, "Efficient Handwritten Character Recognition of MODI Script Using Wavelet Transform and SVD," *in Data Science and Security*, D. S. Jat, S. Shukla, A. Unal, and D. K. Mishra, Eds. Singapore: Springer Singapore, 2021, pp. 227–233, doi: 10.1007/978-981-15-5309-7_24

[21] S. Joseph, A. Datta, O. Anto, S. Philip, and J. George, "OCR System Framework for MODI Scripts using Data Augmentation and Convolutional Neural Network," *in Data Science and Security*, D. S. Jat, S. Shukla, A. Unal, and D. K. Mishra, Eds. Singapore: Springer Singapore, 2021, pp. 201-209, doi: 10.1007/978-981-15-5309-7_21

[22] P. Tamhankar, K. Masalkar, and S. R. Kolhe, "Character Recognition of Offline Handwritten Marathi Documents Written in MODI Script Using Deep Learning Convolutional Neural Network Model," *in Recent Trends in Image Processing and Pattern Recognition, K. C. Santosh and B. Gawali, Eds. Singapore: Springer Singapore*, 2021, pp. 478–487, doi: 10.1007/978-981-16-0507-9_40

[23] K. Mahajan and N. Tajne, "An Ancient Indian Handwritten Script Character Recognition by Using Deep Learning Algorithm," *EFFLATOUNIA: Multidisciplinary Journal*, vol. 5, no. 2, pp. 123-134, Oct. 2021.

[24] S. Chandure and V. Inamdar, "Handwritten MODI Character Recognition Using Transfer Learning with Discriminant Feature Analysis," *IETE Journal of Research*, vol. 69, no. 5, pp. 2584-2594, 2023. doi: 10.1080/03772063.2021.1902867

[25] S. Kulkarni and P. Yannawar, "Recognition of Partial Handwritten MODI Characters Using Zoning," *in Recent Trends in Image Processing and Pattern Recognition, K. C. Santosh and B. Gawali, Eds. Singapore: Springer Singapore*, 2021, pp. 407–430, doi: 10.1007/978-981-16-0507-9_35.

[26] J. Kondhare, V. Yaduvanshi, R. Patil, and R. Kaldate, "Recognition of Handwritten Modi Digits and Characters by Using Deep Learning Algorithm," *International Journal of Emerging Technologies and Innovative Research*, vol. 9, no. 8, pp. 563-572, Aug. 2022. [Online]. Available: http://www.jetir.org/papers/JETIRFP06101.pdf.

[27] M. Ekbote, A. Jadhav, and D. Ambawade, "Implementing a Hybrid Deep Learning Approach to Achieve Classic Handwritten Alphanumeric MODI Recognition," *International Journal of Engineering and Advanced Technology (IJEAT)*, vol. 12, no. 1, pp. 1-8, Oct. 2022, doi: 10.35940/ijeat.A3846.1012122.

[28] V. Pawar, D. Wadkar, S. Kashid, P. Prakare, V. More, and Prof. S. A. Babar, "MODI Lipi Handwritten Character Recognition Using CNN and Data Augmentation Techniques," *International Research Journal of Engineering and Technology (IRJET)*, vol. 09, no. 06, pp. 2345-2351, Jun. 2022.

[29] C. Chandankhede and R. Sachdeo, "Character Recognition using MODI script: Facts, Challenges and its future," *TEST Engineering and Management*, vol. 83, pp. 25389–25395, 2020.

[30] S. Jadhav, V. Inamdar, October 12, 2021, "Handwritten MODI Characters *", IEEE Dataport*, doi: https://dx.doi.org/10.21227/z3gg-8b29.

[31] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA*, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[32] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas,* NV, USA, 2016, pp. 2818-2826, doi: 10.1109/CVPR.2016.308.

Kanchan Varpe

Kanchan Varpe is Research scholar in the Computer Engineering Department at Vishwakarma Institute of Information Technology, Pune, India. Awarded the Master of Engineering Degree from Savitribai Phule Pune University, Sinhgad College of Engineering, Wadgaon, Pune. Research interests include Natural Language Processing (NLP), Image Processing, Computer Networks and Machine Learning.

Dr. Sachin Sakhare

Dr. Sachin R. Sakhare is working as a Professor and Head of the Computer Engineering Department at Vishwakarma Institute of Information Technology, Pune, India. He has 27 Years of experience in engineering education. He is recognized as PhD guide by Savitribai Phule Pune University and currently guiding 8 PhD scholars. He is a life member of CSI, ISTE and IAEngg. He has Published 51 research communications in national, international journals and conferences, with around 393 citations and H-index 7. He has authored 6 books which is published by Springer Nature, CRC Press and IGI Global. He worked as a reviewer of journals published by Elsevier, Wiley, Hindawi, Springer, Inder science, and IETE. He worked as a reviewer for various conferences organized by IEEE, Springer, and ACM. He worked as a member of the Technical and Advisory Committees for various international conferences. Dr. Sachin has Delivered invited talks at various Conferences, FDP's and STTP's as well as to PG and PhD students. He has guided 26 PG students. He has filed and published 07 patents out of which 01 Indian, 03 Australian and 02 south African patents are granted.

Improving Retrieval Performance of Case Based Reasoning Systems by Fuzzy Clustering

F. Saadi[1]*, B. Atmani[2], F. Henni[3]

[1] Laboratoire d'Informatique d'Oran (LIO), University of Oran 1 (Algeria)
[2] Laboratoire d'Informatique d'Oran (LIO), University of Mostaganem (Algeria)
[3] Computer Science and new Technologies Lab (CSTL), University of Mostaganem (Algeria)

* Corresponding author: saadi_fatima@hotmail.fr

Received 20 July 2022 | Accepted 2 June 2023 | Published 5 July 2023

Abstract

Case-based reasoning (CBR), which is a classical reasoning methodology, has been put to use. Its application has allowed significant progress in resolving problems related to the diagnosis, therapy, and prediction of diseases. However, this methodology has shown some complicated problems that must be resolved, including determining a representation form for the case (complexity, uncertainty, and vagueness of medical information), preventing the case base from the infinite growth of generated medical information and selecting the best retrieval technique. These limitations have pushed researchers to think about other ways of solving problems, and we are recently witnessing the integration of CBR with other techniques such as data mining. In this article, we develop a new approach integrating clustering (Fuzzy C-Means (FCM) and K-Means) in the CBR cycle. Clustering is one of the crucial challenges and has been successfully used in many areas to develop innate structures and hidden patterns for data grouping [1]. The objective of the proposed approach is to solve the limitations of CBR and improve it, particularly in the search for similar cases (retrieval step). The approach is tested with the publicly available immunotherapy dataset. The results of the experimentations show that the integration of the FCM algorithm in the retrieval step reduces the search space (the large volume of information), resolves the problem of the vagueness of medical information, speeds up the calculation and response time, and increases the search efficiency, which further improves the performance of the retrieval step and, consequently, the CBR system.

Keywords

Case Based Reasoning, Case Retrieval, Classification, Data Mining, Decision Support, Fuzzy Clustering, Immunitherapy, Kmeans.

DOI: 10.9781/ijimai.2023.07.002

I. Introduction

Case Based Reasoning (CBR) is a problem-solving paradigm. Pantic [2] and Aamodt & Plaza [3] defined a reasoning cycle with four phases: Retrieve-Reuse-Revise-Retain. Instead of relying only on general knowledge of a problem domain, CBR relies on the retrieval of past and solved problems, called source cases, to solve a current problem, called a target problem. A new experience is maintained each time a problem has been solved, making it immediately available for future problems. That is why the retrieval of similar cases is a crucial phase in the CBR cycle.

Dependent on the process of medical situation resolution, it is clear that the doctor's reasoning is mostly based on the fact that the current situation is probably treated before, and so the doctor will propose a solution that is more or less identical to the one previously adopted. This reasoning resembles the CBR reasoning methodology. This has motivated a lot of research into this reasoning method in the medical field [4], leading to the creation of computerized tools for solving decision-making problems using only this reasoning method (CBR). This work has ramifications in various fields of artificial intelligence: knowledge representation, classification, similarity measures, etc., which has made it a complex but widely used reasoning mode in medical decision support. However, The application of classical CBR systems in the medical domain has limitations due to the increasing complexity of this domain since many healthcare applications are simply too complex and multifaceted to be processed using this methodology [5]. The case representation has become more complex than in history in several applications: medical information (case) may come in part in the format of time series, images, or free text, and may also be intrinsically high-dimensional, imprecision, vagueness, and uncertainties [6]. Indeed, the large volume of generated medical information (symptoms, diseases, and treatments) which is increasing slows down the similarity computation in the retrieval phase and it becomes very expensive in computation time, knowing that time is a very important factor not to be neglected in a medical diagnosis [7]. Therefore, a suitable retrieval algorithm must be chosen to solve these problems [8].

Please cite this article as:
F. Saadi, B. Atmani, F. Henni. Improving Retrieval Performance of Case Based Reasoning Systems by Fuzzy Clustering, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 1, pp. 84-91, 2024, http://dx.doi.org/10.9781/ijimai.2023.07.002

The richness of various reasoning methods or approaches may be integrated to exceed the limitations of the application of classical CBR and support it as a knowledge engineering methodology at each phase of its cycle [9]. This integration has been widely deployed in multi-modal reasoning systems and it has been shown to be well-adapted, in particular for work related to the medical domain [8]. Among the techniques integrated into the classical CBR, data mining methods have shown some advantages in particular for improving the retrieval step [10].

The objective of this work, which is an extension of research presented in [11], is to improve the CBR cycle by improving the performance of the most important step in the CBR cycle, retrieve, through the choice of the best similarity measure. The proposed approach is adopted to develop a CDSS that assists dermatology experts in predicting a patient is responding to immunotherapy therapy for warts.

In this paper, we propose a new approach that integrates one of the data mining techniques which is clustering (Fuzzy C-Means (FCM) and K-Means) in the CBR cycle. The integration of clustering aims to demonstrate the value of this technique in CBR to reduce the number of cases while reducing the complexity of the retrieval step and speeding up the time which is a very important aspect of medical diagnosis. Thus, to prove the purpose of fuzzy logic to model the vague, imprecise, and uncertain concepts of medical information, we chose to integrate the FCM technique in the retrieval step. Fuzzy CBR should be used for this reason, as well as for better decision support [12]. Section II in this paper focuses on some related works. The methodologies provided in this study are explained in Section III. Experiments are implemented in Section IV, and the results are interpreted and evaluated. Finally, Section V, ends with conclusion.

II. Background

A. Use of CBR

CBR is a general decision-making methodology used in the medical field [7]. Several studies have used CBR in this area. Sharma and Mehrotra [13] applied the retrieval of cases by similarity measurement to develop a CBR implementation for the diagnostic of chronic renal failure. Demigha [14] designed a generic eLearning application for radiologists and other hospital staff. They developed this instrument using the CBR method. Mansoul and Atmani [10] proposed using Multi-Criteria Analysis (MCA) with the standard CBR retrieval to aid in finding the ideal solution. They included this method in a clinical decision support system. Gu et al. [15] implemented a CBR system for breast cancer diagnosis and used it in two experiments, one on benign/malignant tumor prediction and the other on secondary cancer prediction. Benfriha et al. [16] developed a new approach for case acquisition in CBR based on multi-label text categorization applied in a child's traumatic brain injuries dataset. El-Sappagh et al. [17] demonstrated that non-clinical CBR systems have made more progress than clinical CBR systems. In addition, when contrasted to other diabetes healthcare systems, CBR systems achieve the smallest gains. These studies show that clinical CBR, especially in diabetic systems, requires more thorough improvements.

B. Use of Data Mining

Data mining was crucial in the development of intelligent healthcare systems [18], [19]. We provide a list of research papers that employ data mining techniques in the healthcare industry. Dewan Sharma [20] created a tool that can identify and retrieve new heart illness information from a previous heart disease database record. They applied data mining algorithms including Neural Networks, Decision Tree, and Naive Bayes for their proposal. The idea of this research is to handle complex requests in diagnosing heart disease, allowing health doctors to improve medical judgment more than standard decision support systems could. Chen et al. [21] conducted research on localized chronic cerebral infarction and presented a new illness prediction method that has multiple modes and based on convolutional neural networks. Kumar & Sahoo [22] introduced a novel approach that uses Naive Bayes as well as genetic algorithms to enhance cardiovascular disease prediction. The outcomes of their research demonstrate that this approach enhances the efficiency of heart disease detection. Chaurasia & Pal [23] developed prediction models for heart disease survivability using a large dataset and applied three data mining techniques including decision trees and rule-based classifiers. Hachesu et al. [24] presented a method for determining and predicting heart patient length of hospital stay, they adopted in this research the decision trees, Support Vector Machines, and Artificial Neural Networks data mining algorithms. Martín et al. [25] created An algorithm for semi-supervised clustering. The technique, which is based on an ensemble of dissimilarities, has been used to identify tumor samples using gene expression patterns.

Clustering is a crucial unsupervised data mining technique used to find some underlying structure in a collection of patterns or objects [26]. A cluster maximizes the similarity of these objects and minimizes the similarity of objects not belonging to it. To do this, the data mining process uses distance functions. These functions evaluate the existing similarities (distances) between the entities to be grouped. In order to uncover the dataset's underlying natural cluster patterns, the choice of similarity measure is crucial [1]. Many distance functions are available in the literature. Saadi et al. In this work, we adopted the standard version of K-Means and FCM that uses the Euclidean distance, but there are many works that improve these techniques by using other distances. Kapil & Chawla [27] used Manhattan distance with C-means clustering. Sharma et al. [28] integrated the S-distance and the Euclidean distance with the C-means clustering algorithm. Karlekar et al. [26] added the S-distance to the traditional fuzzy K-means method in place of the Euclidean distance. Seal et al. [29] developed Fuzzy c-means clustering using a novel similarity metric based on Jeffreys-divergence.

1. Combining CBR With Data Mining Techniques

Some works that integrate data mining techniques in each step of the CBR cycle has illustrated in Table I.

TABLE I. Medical Systems Integrated Data Mining Techniques in CBR Cycle

References	CBR process and methods	Application domain
[30]	Retrieve: Euclidean, Manhattan, or Hamming distance Adaptation decision rules	medical
[31]	Retrieve: KNN Adaptation ANN	representations of human organs
[32]	Retrieve: KNN Adaptation: Rule-Based Reasoning	Cancer diagnosis
[23]	Retrieve: Dissimilarity Measurementsrevises and reuses genetic algorithm	Medical Diagnosis
[21]	Retrieve: Decision treeAdaptation: Decision Rule extracted from Decision tree	cardiovascular disease
[33]	case acquisition: Multi-label Retrieve: + KNN	child's traumatic brain injuries

Retrieve is often regarded as the most important step in CBR systems. The similarity measurement is the main task of this step. From the case base, the process will select similar cases that will be deemed the most relevant to begin the process of determining the solution for the medical situation. Several research papers have looked at the use of data mining algorithms to improve the efficiency of the retrieval stage. Gu et al. [34] and Jung et al. [35] proposed CBR-based models which combined the naive Bayes and KNN approaches for similarity measurement. Benbelkacem et al. [31], Chen et al. [21], and Saadi et al. [36] integrated the decision trees and the KNN in the retrieval step for the similarity calculation. Mansoul & Atmani [7] and Khussainova & Jagannathan [37], Koo et al. [38] and Yadav [39] improved the retrieval step by decreasing the search space using Clustering techniques. In faced with complex real-world applications, retrieving cases must deal with uncertainties [30]. Demigha [14] focused at the function of the fuzzy system in the various stages of CBR and found that integrating fuzzy logic with CBR resulted in efficient hybrid approaches. Geetha et al. [40] presented a fuzzy CBR strategy for deciding the urgency of COVID-19 sick people. Ibrahim & Odedele [41] developed a system for detecting and diagnosing infectious diseases, COVID-19, using fuzzy CBR. Choudhury et al. [42] used fuzzy-rough nearest neighbor to enhance the retrieval step of the CBR system's performance and efficiency. The experimental findings reveal that this combination beats KNN for classification by a large margin, effectively boosting case retrieval efficiency and performance. Yamin et al. [43] offered a case-matching process that uses two algorithms to find similar cases the case similarity algorithm in the case when the case database is small and FCM secondary retrieval algorithm in the case when the case base is large. Banerjee & Chowdhury [44] used the (FCM) algorithm in the CBR system to classify the most prevalent anomalies in retina images caused by maturity-level eye disease and diabetes. Ekong et al. [6] developed a clinical decision support system based on CBR, neural networks, and fuzzy logic for diagnosing depressive illnesses. Begum et al. [30] created a fuzzy CBR approach that categorizes healthy and distressed people. Benamina et al. [12] developed a fuzzy CBR technique to ensure a better diagnosis for diabetics, which includes fuzzy-decision trees with CBR to increase reaction speed and recovery accuracy of similar cases. In previous work [11], a medical decision support system has been proposed; this system is guided by case reasoning and the clustering technique. The research aimed to enhance the retrieving process by integrating the FCM method in the similarity calculation. The results of the experiments performed in this paper were encouraging since they improved the most critical phase in the CBR cycle (retrieve). In this work, we propose an extended version of the article proposed in [11].

III. Contribution

The authors of this research develop a new approach that integrates the clustering algorithms(K-Means and FCM) in the retrieval step. this approach aims to speed up the similarity calculation which accelerates the retrieval phase and improves its performance. Fig. 1 summarizes the essential phases of the proposed approach. Our approach's primary steps are as follows:

- Identify the number of clusters using the elbow approach, then apply K-means techniques to create the clusters;
- Define the membership degree matrix using the FCM algorithm;
- Using the jCOLIBRI platform, construct the base case;
- The start of the CBR cycle begins with the arrival of a new case by launching the retrieval phase. In this phase, the user chooses the K-means or FCM for the retrieval step:

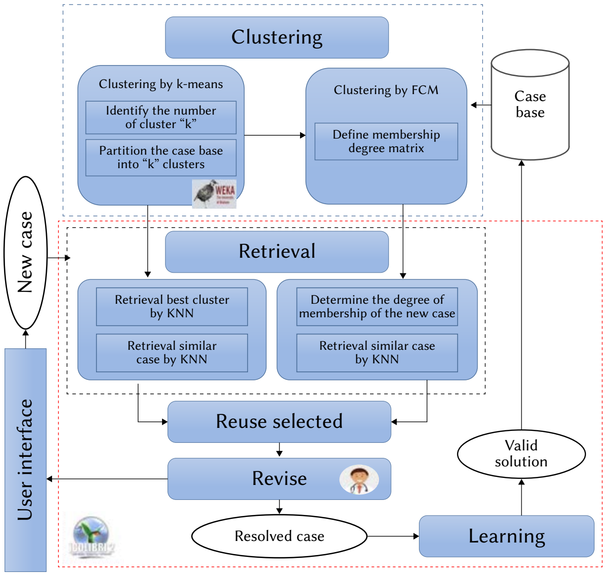

Clustering
Clustering by k-means
Identify the number of cluster "k"
Partition the case base into "k" clusters
WEKA
Clustering by FCM
Define membership degree matrix
Case base
New case
Retrieval
Retrieval best cluster by KNN
Retrieval similar case by KNN
Determine the degree of membership of the new case
Retrieval similar case by KNN
Reuse selected
Revise
Valid solution
User interface
Resolved case
Learning
jCOLIBRI

Fig. 1. The Proposed approach's architecture.

- Retrieval with K-means: retrieval of the best cluster and similar cases with the KNN algorithm.
- Retrieval with FCM: To identify the ideal clusters where the KNN algorithm technique measures the similarity, determine the new case's membership degree in each cluster.

A. Clustering Techniques

These techniques are used as part of a solution-finding strategy that helps you to pick the optimal answer from a smaller number of options. We chose the K-means and FCM as unsupervised classification techniques for clustering, with the goal of structuring the case base, guiding, and speeding up the retrieval.

1. Clustering With K-means

The method consists in splitting the data into k clusters. It starts with a random clustering of the data (into k clusters), then assigning every element to the cluster that is nearest to it. Once the first iteration is completed, the averages of the clusters are calculated and the process is repeated until the clusters are stabilized. In this work, the clusters were generated from the data presented in a CSV file using the K-means algorithm implemented in the WEKA platform [45]. This algorithm has a fundamental drawback in that it requires the number of clusters, K, to be provided [46]. A good classification produces classes with a strong similarity within each class and a minor similarity between different classes. The distance between a point and its cluster center is called intra-class inertia. It's will be quantified as in (1) [46].

$$Intra-class = \sum_{i=1}^{k} \sum_{x \epsilon C_i} \parallel x - z_i \parallel^2 \tag{1}$$

Where k represents the total number of clusters, and C_i represents the cluster center. The distance between the clusters is the inter-class inertia. The distance between cluster centers is calculated, and the lowest of these numbers is utilized to determine it. The equation of inter-class inertia is defined in (2) [46].

$$Inter-class = min(\parallel z_i - z_j \parallel^2), i = 1,2,\ldots,k; j = 1,2,\ldots.,k \tag{2}$$

The partition is excellent when the classes are homogeneous, the intra-class inertia is low, and the inter-class inertia is high. One of the existing strategies for identifying the clusters number is the elbow method. It is a visual method of varying the number of clusters and monitoring the evolution of a solution quality indicator (the proportion of inertia) in order to search for the "elbow" in the graphic. The idea is, to begin with $K = 2$, and keep increasing it in each step by 1, calculating the clusters and the intra-class inertia. The cost drops dramatically at some value for K and afterward, it reaches a plateau when you increase it further. This is the value we want for K. [47]. The K-means technique is applied to split the case base into K clusters after selecting the best value of K. As a result, we get a 0 and 1 Boolean matrix. The existence of cases in the cluster is indicated by a 1, whereas a 0 indicates the lack of this case.

2. Clustering With Fuzzy C-Means (FCM)

The FCM method, sometimes called the fuzzy K-means, was proposed by [48]. It is a fuzzy logic method in which the probability of belonging to the c groups for each observation is evaluated and presented as a membership matrix u of size n by c, where u_{ik} is the probability of observation k belonging to group i. The total of each row u equals one, like K-means, requires that the number of classes and iterations be fixed in advance; the initial class centers are also either randomly drawn or specified by the user. The class centers v are then iteratively updated with the matrix u. The membership probabilities are then re-evaluated using the new class centers. This method is performed till the given number of iterations has been attained or a convergence criterion is met. (3) is used to calculate the update of the classes and u:

$$u_{ik} = \frac{(\| x_k - v_i \overset{2}{\|})^{\frac{-1}{(m-1)}}}{\sum_{j=1}^{c}(\| x_k - v_j \overset{2}{\|})^{\frac{-1}{(m-1)}}} et v_i = \frac{\sum_{k=1}^{N} u_{ik}^{m} (x_k)}{\sum_{k=1}^{N} u_{ik}^{m}} \tag{3}$$

Where k and i represent an observation and a class, respectively. The main addition to the FCM method is the parameter m. It allows controlling the degree of fuzziness in the classification. If $m = 1$, the obtained classification is strict (the matrix u contains only 0 and 1 values); as m increases, the values of the matrix u decrease until they are perfectly homogeneous. Different types of distances (Euclidean, Chi-square, Mahalanobis, etc.) can be used as in the classical K-means, and the initial class centers, as well as the number of iterations, can potentially affect the classification results. It is common practice to experiment with various values for the number of classes until the ideal combination is found (More information, including an overview of the FCM algorithm, can be found in our previous work [11]. We applied the FCM approach in this step. The membership degree matrix determined from the K-means algorithm is the input. We get a membership degree matrix for each instance in each cluster as a result of this step.

B. CBR Process

1. Retrieval With K-Means Method

The jCOLIBRI platform is used for the retrieval stage. Once a new case is received, similar cases are chosen using the KNN algorithm to calculate similarity. Our search for similar cases is divided into two stages:

- The best cluster search: In this step, the similarity calculation is applied between the target case and the cluster centers (centroids) found via the K-means method.
- The search for similar cases: This step allows to filter the cases so that only the cases belonging to the best cluster are kept and that the similarity computation will be performed between the selected and filtered cases, instead of being computed between all of the case base's cases, which facilitates the retrieval step.

2. Retrieval With FCM Method

In this step, the FCM method is relaunched to calculate the membership degree of the new case in the cluster and takes from the resulting matrix the two clusters to which the new case has a high degree of membership and that is the advantage over the K-means. It is up to say instead of searching in one and only one cluster, with FCM, the search will be done in more than one cluster and this gives a high percentage to finding the best case similar to the new problem. The similarity is calculated only between the cases of the chosen clusters in the previous step and the KNN is used to select similar cases. To calculate the local similarity (similarity between attributes), the Euclidean distance is used because we only have attributes of numeric type; the formula of the Euclidean distance is defined by(4):

$$D(x, y) = \sqrt{\frac{x^2 - y^2}{x + y}} \tag{4}$$

After the computation of the local similarity, the global similarity is measured (the distance between two cases) using the technique most used in CBR systems KNN. The following equation is used to calculate similarity by KNN as in (5):

$$sim(C, S) = \frac{\sum_{f=1}^{n} w_f * sim(C_s, S_f)}{\sum(w_f)} \tag{5}$$

C stands for a new case, S for a saved case, w for an expert-defined weight, n represented case's attributes number, f for the attribute index and $Sim(C, S)$ is the local similarity for attribute f. Adaptation is not taken into account in our study. Because by definition it allows to partially or completely resume the solution that already exists in the database which is not our case. The revision entails validating the solution devised by the expert -doctor). The solution to the new problem has been discovered and validated, a new experiment is created, and it is saved in the case database to expand the case database and boost the potential for solving future situations.

IV. Experimentation and Results

A. Construction of the Case Base

The UCI Machine Learning Repository is used for evaluating the performance and efficiency of the proposed system. The authors specifically selected the "Immunotherapy" dataset which will be discussed further in the following. The UCI "Immunotherapy" dataset https://archive.ics.uci.edu/ml/datasets/Immunotherapy+Dataset has been obtained inside a clinical of dermatology in Mashhad city situated in Iran. It includes 90 instances that presented the patients who suffered from warts. Each instance consists of 08 attributes, as shown in Table II. Such as the class that signifies the patients' response to immunotherapy treatment (failure or success).

TABLE II. Features Used in the Immunotherapy Dataset

N°	Attribute	Value
1	Gender(Sex)	Man or Woman
2	Age	From 15 to 56 years
3	Time passed before therapy (Time)	From 0 to 12 months
4	The amount of warts on the body (N° warts)	From 1 to 19 warts
5	Wart's type (Type)	common, plantar, or both
6	The biggest wart's surface area (Area)	From 6 to 900 mm2
7	The initial test's induration diameter (Ind- dia)	From 5 to 70 mm
8	The patients' response to treatment (Class)	Success or failure

B. Clustering Technique

1. Clustering With K-Means

In the beginning, the approach generates the clusters using the K-means algorithm implemented under the WEKA platform. To determine the cluster number, the elbow method is appliqued: we vary k and follow the intra-class inertia, Fig. 2 illustrates the variation in intra-class inertia as a function of the number of clusters selected. According to this result, we deduce that the partition in $K = 6$ is the last to induce a significant information gain (the curvature in Fig. 2 shows a clear peak for $K = 6$ clusters). The K-means algorithm restart and generates six clusters.

2. Clustering With FCM Method

The system imports the case base functionality and starts configuration after it has been launched. Furthermore, it builds the membership degree matrix using the FCM technique.

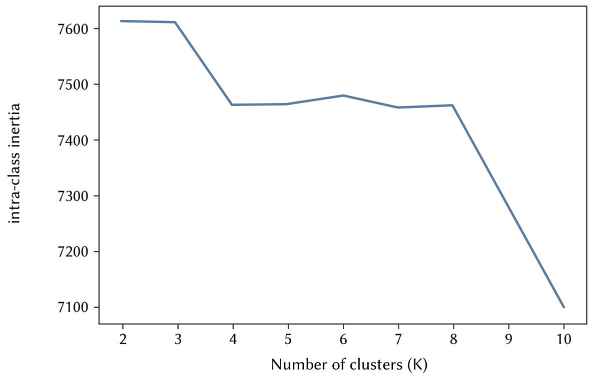


Fig. 2. The Intra-class inertia's variation.

3. Case Based Reasoning

The system consists of a simple interface via which a request may be launched. This request definitely illustrates a new patient with warts. The interface also provides the clinician the opportunity of specifying the amount of similar cases that he wants to retrieve. This option allows the doctor to define the number of cases to retrieve.

4. Retrieval With K-Means Method

The system calculates the similarity between centroids cases (cluster centers) and finds the best cluster that groups the nearest cases to the new case, then it filters the cases from the base so that only the best cluster cases are kept and it restarts the similarity calculation between the filtered cases to find the most similar cases requested.

5. Retrieval With FCM Method

The system moves on to the retrieval stage, where it re-runs the FCM technique to construct the two clusters in which the target case has a high degree of membership and search for the best similar case inside these two clusters rather than researching the whole case base. The clinician can go through similar cases and choose the one that he thinks is the most effective. The user can then make adjustments to the obtained solution after making this choice. Once the doctor has adjusted the solution, the system gives the user the option of retaining this target case and saving it with the cases of the case base or putting it in a temporary case base, where it may require more time before choosing whether to keep the target case.

C. Performance Evaluation

To evaluate the retrieval with FCM and K-means methods, we assessed them on a similar case base. These approaches' goal is to predict a patient with warts (target case) is responding to immunotherapy therapy (solution for the target case). In the following table Table III, five of target cases are randomly chosen from the cases from the case base, considering that the entire case base is used for the generation of the models. Then we execute the clustering using both methods and the retrieval using KNN ($K = 3$).

Table IV shows the results of experimentation obtained by two methods. Retrieving with K-means returns the best cluster, whereas retrieving with FCM provides the two best clusters with the of the target case's membership degree to these clusters, and the highest similar cases with the degrees of similarity and response of the target case to treatment.

TABLE III. Target Cases

Case	Sex	Age	Time	N° warts	Type	Area	Ind-dia	Y(CLASS)
Case 10	Women	32	12	6	3	35	5	Failure
Case 6	Man	15	5	3	3	84	7	Success
Case 81	Man	23	3	2	3	87	70	Success
Case 32	Man	30	1	2	1	88	3	Success
Case 89	Man	32	12	9	1	43	50	Failure

TABLE IV. Results Obtained by Retrieving With K-means and FCM

	Retrieving with K-means				Retrieving with FCM				
New case	No. cluster	No. Similar case	Similarity degree	Response to treatment	No. cluster	Membership degree	No. Similar case	Similarity degree	Response to treatment
Case N°10	1	Case23	0.32	Failure	5	0.28	Case 10	0.0	Failure
		Case 90	0.40	Success	2		Case 17	0.29	Success
		Case 11	0.41	Success			Case 45	0.37	Success
Case N° 6	2	Case6	0.00	Success	4	0.29	Case 6	0.0	Success
		Case 72	0.20	Success	3		Case 66	0.22	Success
		Case 66	0.22	Success			Case 36	0.23	Failure
Case N°81	2	Case21	0.34	Success	4	0.32	Case 81	0.0	Success
		Case 32	0.40	Success	3		Case 71	0.35	Success
		Case 6	0.47	Success			Case 42	0.34	Success
Case N°32		Case61	0.51	Success	4	0.28	Case 32	0.0	Success
		Case 63	0.58	Success	3		Case 4	0.25	Failure
		Case 74	0.59	Success			Case 6	0.27	Success
Case N°89		Case10	0.42	Failure	5	0.26	Case 89	0.0	Failure
		Case 27	0.45	Success	2		Case 21	0.25	Failure
		Case 18	0.45	Success			Case 59	0.27	Success

The quality of a diagnosis is crucial in providing medical treatment since a doctor's recommendations for medical therapy are based on diagnostic tests (medical tests, medical signs, symptoms, etc.). Fortunately, it is possible to measure the features of diagnostic tests. Based on these features, the ideal test may be selected for a particular illness condition. A diagnostic test is frequently described using the statistics of sensitivity, specificity, and accuracy. They are used in particular to measure a test's quality and dependability [49]. Several parameters are frequently used in conjunction with the definitions of accuracy, sensitivity, and specificity. True positives *TP* , false positives *FP* , true negatives *TN*, and false negatives *FN* are the four parameters of the confusion matrix (Table V.). The outcome of the diagnostic test is regarded as a true positive if a disease is demonstrated to be present in a patient and the diagnostic test also demonstrates the existence of the disease. Similar to this, when a disease is absent in a patient and the diagnostic test indicates this is also the case, the test result is said to be a true negative (*TN*). True results, whether positive or negative, point to a correlation between the outcome of the diagnostic test and the established condition (also called the standard of truth). No clinical exam, though, is flawless. The test result is considered to be false positive if it shows that a patient has an illness when they actually don't (*FP*). Similar to this, a test result is false negative if it indicates that a patient who has a condition for sure does not have it (*FN*). The test findings are at odds with the real disease when they are both falsely positive and falsely negative.

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \tag{6}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{7}$$

$$Specificity = \frac{TN}{TN + FP} \tag{8}$$

According to the equations (6),(7),(8), sensitivity is the percentage of true positives that a diagnostic test successfully identifies. It demonstrates how accurate the test is in identifying diseases. The percentage of true negatives that a diagnostic test accurately identifies is known as specificity. It demonstrates how well the test detects a normal (negative) situation. Accuracy is the percentage of real outcomes in a population, whether they are real positive or real negative. It gauges how reliable a condition-specific diagnostic test is. to assess the efficacy of our system, these performance measurements are calculated in Table IV.

TABLE V. Confusion Matrix

	Predicted Class	
Actual class	True Positives (TP)	False Positives (FP)
	False Negatives (FN)	True Negatives (TN)

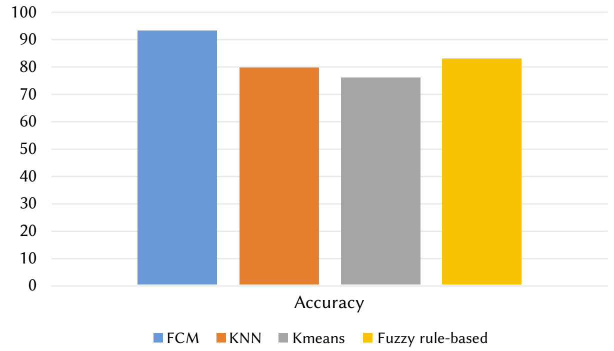


Fig. 3. Comparison of results of experimentation

In previous work [36], we applied KNN in the retrieval phase, we also tested K-means in the same way in order to compare the results, while [50] applied Fuzzy rule-based on the same case base that we used and for the same objective as ours, which is the prediction of the result of immunotherapy treatment.

In Fig. 3, We present the results of a comparison between our proposed approach FCM and other methods (KNN, Kmeans, and Fuzzy rule-based).

D. Discussion of Results

As shown in Table IV, the retrieval step with FCM technique was always successful in finding the cluster containing the target case, however, the retrieval with K-means was only successful in finding the cluster containing case n°6. Thus, the similarity degrees provided by retrieval by FCM are higher than those of retrieval by k-means. We also observe that both techniques provide give the right result of treatment for all the target cases. This confirms that the integration of FCM clustering in the retrieval step has succeeded in improving the predictive accuracy. Table VI confirms these results whereas we observed in Fig 3, FCM retrieval offers good accuracy (93.33), sensitivity (92.59), and specificity (100). Even when comparing these results with other approaches, the proposed approach obtains the highest accuracy (93.33) compared to other techniques.

TABLE VI. Performance Evaluation

Accuracy	93.33%
Sensitivity	92.59%
Specificity	100%

In conclusion, the encouraging results obtained show that the integration of FCM clustering in the CBR cycle, precisely in the retrieval stage, allows the achieving of better performances and accelerates the similarity computation by reducing the search space, which leads to accelerated search time, improves the retrieval stage and consequently the CBR system, solves the problems of using the classical CBR system in the medical domain such as the large volume of generated medical data and the complexity and uncertainty of these data, and finally leads to a better medical decision making.

V. Conclusion

In this paper, we primarily offer a new approach that integrated the Clustering techniques (K-means and FCM) in the CBR cycle. This integration aims to enhance the retrieval step and consequently the CBR system in order to resolve the problems with applying the classical CBR system in the medical field. The proposed approach has been applied to the immunotherapy dataset in order to predict the response of patients with warts to the immunotherapy.

Experiments have demonstrated that the strategy based on CBR and fuzzy clustering (FCM) was successful in improving the performance of retrieval step such as the accuracy, case retrieval precision, and calculation time. It was discovered that this approach may greatly and effectively minimize the number of cases (research space), solve the problem of the complexity, and the uncertainty of medical information, speed up the similarity calculus, and increase the effectiveness of the search, allowing us to achieve our goal and resolve the problem of classical CBR and improve it.

This approach uses the standard version of k-means and FCM and adopts the Euclidean distance to generate clusters. However, there are several improved versions of these techniques such as [26], [1], [28], [29],... and they have achieved successful results. As a future work we aim to improve our approach by improving the similarity measures (distances) used in the clustering.

References

[1] A. Seal, E. Herrera Viedma, et al., "Performance and convergence analysis of modified c-means using jeffreys-divergence for clustering," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 2, pp. 141–149, 2021, doi: https://doi.org/10.9781/ijimai.2021.04.009.

[2] A. Aamodt, E. Plaza, "Case-based reasoning: Foundational issues, methodological variations, and system approaches," *AI communications*, vol. 7, no. 1, pp. 39–59, 1994.

[3] M. Pantic, "Introduction to machine learning & case- based reasoning," *London: Imperial College*, 2005.

[4] N. Choudhury, S. A. Begum, "A survey on case- based reasoning in medicine," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 8, 2016.

[5] S. Montani, "Exploring new roles for case-based reasoning in heterogeneous ai systems for medical decision support," *Applied Intelligence*, vol. 28, no. 3, pp. 275–285, 2008.

[6] V. E. Ekong, U. G. Inyang, E. A. Onibere, "Intelligent decision support system for depression diagnosis based on neuro-fuzzy-cbr hybrid," *Modern Applied Science*, vol. 6, no. 7, p. 79, 2012.

[7] A. Mansoul, B. Atmani, "Clustering to enhance case- based reasoning," in *Modelling and Implementation of Complex Systems*, Springer, 2016, pp. 137–151.

[8] R. Schmidt, L. Gierl, "Prognostic model for early warning of threatening influenza waves," in *1st German workshop on experience management: sharing experiences about the sharing of experience*, 2002, Gesellschaft für Informatik eV.

[9] I. Bichindaritz, C. Marling, "Case-based reasoning in the health sciences: Foundations and research directions," in *Computational Intelligence in Healthcare 4*, Springer, 2010, pp. 127–157.

[10] M. Abdelhak, A. Baghdad, "Combining multi-criteria analysis with cbr for medical decision support," *Journal of Information Processing Systems*, vol. 13, no. 6, pp. 1496– 1515, 2017.

[11] F. Saadi, B. Atmani, F. Henni, "Integration of fuzzy clustering into the case base reasoning for the prediction of response to immunotherapy treatment," in *Mediterranean Conference on Pattern Recognition and Artificial Intelligence*, 2019, pp. 192–206, Springer.

[12] M. Benamina, B. Atmani, S. Benbelkacem, "Diabetes diagnosis by case-based reasoning and fuzzy logic," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 3, pp. 72–80, 2018.

[13] S. Sharma, D. Mehrotra, "Building cbr based diagnosis system using jcolibri," in *2017 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence*, 2017, pp. 634–638, IEEE.

[14] S. Demigha, "A generic elearning tool for radiologists and hospital practitioners with cbr," in *European Conference on e-Learning*, 2015, p. 809, Academic Conferences International Limited.

[15] D. Gu, C. Liang, H. Zhao, "A case-based reasoning system based on weighted heterogeneous value distance metric for breast cancer diagnosis," *Artificial intelligence in medicine*, vol. 77, pp. 31–47, 2017.

[16] H. Benfriha, B. Atmani, B. Khemliche, N. T. Aoul, Douah, "A multi-labels text categorization framework for cerebral lesion's identification," in *International Conference on Computing*, 2019, pp. 103– 114, Springer.

[17] S. El-Sappagh, M. M. Elmogy, "Medical case based reasoning frameworks: Current developments and future directions," *Virtual and Mobile Healthcare: Breakthroughs in Research and Practice*, pp. 516–552, 2020.

[18] C. Aflori, M. Craus, "Grid implementation of the apriori algorithm," *Advances in engineering software*, vol. 38, no. 5, pp. 295–300, 2007.

[19] K. Srinivas, G. R. Rao, A. Govardhan, "Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques," in *2010 5th International Conference on Computer Science & Education*, 2010, pp. 1344–1349, IEEE.

[20] A. Dewan, M. Sharma, "Prediction of heart disease using a hybrid technique in data mining classification," in *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)*, 2015, pp. 704–706, IEEE.

[21] M. Chen, Y. Hao, K. Hwang, L. Wang, L. Wang, "Disease prediction by machine learning over big data from healthcare communities," *Ieee Access*, vol. 5, pp. 8869–8879, 2017.

[22] S. Kumar, G. Sahoo, "Classification of heart disease using naive bayes and genetic algorithm," in *Computational Intelligence in Data Mining-Volume 2*, Springer, 2015, pp. 269–282.

[23] V. Chaurasia, S. Pal, "Early prediction of heart diseases using data mining techniques," *Caribbean Journal of Science and Technology*, vol. 1, pp. 208–217, 2013.

[24] P. R. Hachesu, M. Ahmadi, S. Alizadeh, F. Sadoughi, "Use of data mining techniques to determine and predict length of stay of cardiac patients," *Healthcare informatics research*, vol. 19, no. 2, pp. 121–129, 2013.

[25] M. Martín Merino, A. J. López Rivero, V. Alonso, M. Vallejo, A. Ferreras, "A clustering algorithm based on an ensemble of dissimilarities: An application in the bioinformatics domain.," *International Journal of Interactive Multimedia & Artificial Intelligence*, vol. 7, no. 6, 2022.

[26] A. Karlekar, A. Seal, O. Krejcar, C. Gonzalo-Martin, "Fuzzy k-means using non-linear s-distance," *IEEE Access*, vol. 7, pp. 55121–55131, 2019.

[27] S. Kapil, M. Chawla, "Performance evaluation of k-means clustering algorithm with various distance metrics," in *2016 IEEE 1st international conference on power electronics, intelligent control and energy systems (ICPEICES)*, 2016, pp. 1–4, IEEE.

[28] K. K. Sharma, A. Seal, "Clustering analysis using an adaptive fused distance," *Engineering Applications of Artificial Intelligence*, vol. 96, p. 103928, 2020.

[29] A. Seal, A. Karlekar, O. Krejcar, C. Gonzalo-Martin, "Fuzzy c-means clustering using jeffreys-divergence based similarity measure," *Applied Soft Computing*, vol. 88, p. 106016, 2020.

[30] S. Begum, M. U. Ahmed, S. Barua, "Multi-scale entropy analysis and case-based reasoning to classify physiological sensor signals," *Luc Lamontagne and Juan A. Recio-Garcıa (Editors)*, vol. 129, 2012.

[31] S. Benbelkacem, B. Atmani, M. Benamina, "Treatment tuberculosis retrieval using decision tree," in *2013 international conference on control, decision and information technologies (CoDIT)*, 2013, pp. 283–288, IEEE.

[32] X. Blanco, S. Rodríguez, J. M. Corchado, C. Zato, "Case-based reasoning applied to medical diagnosis and treatment," in *distributed computing and artificial intelligence*, Springer, 2013, pp. 137–146.

[33] H. Benfriha, B. Atmani, F. Barigou, F. Henni, B. Khemliche, S. Fatima, A. Douah, Z. Z. Addou, "Improving cbr retrieval process through multilabel text categorization for health care of childhood traumatic brain injuries in road accident," in *Proceedings of Sixth International Congress on Information and Communication Technology, 2022, pp. 721–731,* Springer.

[34] D. Gu, W. Zhao, Y. Xie, X. Wang, K. Su, O. V. Zolotarev, "A personalized medical decision support system based on explainable machine learning algorithms and ecc features: Data from the real world," *Diagnostics*, vol. 11, no. 9, p. 1677, 2021.

[35] Y.-G. Jung, B. Kim, H. Nam, M. Rhee, J.-S. Lee, "Effective diagnosis of coronary artery disease using case-based reasoning," *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, no. 5, pp. 449–457, 2021.

[36] F. Saadi, B. Atmani, F. Henni, "Integration of datamining techniques into the cbr cycle to predict the result of immunotherapy treatment," in *2019 International Conference on Computer and Information Sciences (ICCIS)*, 2019, pp. 1–5, IEEE.

[37] G. Khussainova, S. Petrovic, R. Jagannathan, "Retrieval with clustering in a case-based reasoning system for radiotherapy treatment planning," in *Journal of Physics: Conference Series*, vol. 616, 2015, p. 012013, IOP Publishing.

[38] C. Koo, W. Li, S. H. Cha, S. Zhang, "A novel estimation approach for the solar radiation potential with its complex spatial pattern via machine-learning techniques," *Renewable energy*, vol. 133, pp. 575–592, 2019.

[39] P. Yadav, "Case retrieval algorithm using similarity measure and adaptive fractional brain storm optimization for health informaticians," *Arabian Journal for Science and Engineering*, vol. 41, no. 3, pp. 829–840, 2016.

[40] S. Geetha, S. Narayanamoorthy, T. Manirathinam, D. Kang, "Fuzzy case-based reasoning approach for finding covid-19 patients priority in hospitals at source shortage period," *Expert Systems with Applications,* vol. 178, p. 114997, 2021.

[41] H. D. Ibrahim, T. O. Odedele, "Covid19 infectious disease detection and diagnosis system using case- based reasoning and fuzzy logic inference model," in *International Conference on Intelligent and Fuzzy Systems*, 2021, pp. 162–170, Springer.

[42] N. Choudhury, S. A. Begum, "Neuro-fuzzy-rough classification for improving efficiency and performance in case-based reasoning retrieval,"

in *Computational Network Application Tools for Performance Management*, Springer, 2020, pp. 29–38.

[43] Z. Yamin, Z. Mengmeng, G. Xiaomin, Z. Zhiwei, Z. Jianhua, "Research on matching method for case retrieval process in cbr based on fcm," *Procedia engineering, vol. 174, pp. 267–274, 2017.*

[44] S. Banerjee, A. R. Chowdhury, "Case based reasoning in the detection of retinal abnormalities using decision trees," *Procedia Computer Science*, vol. 46, pp. 402–408, 2015.

[45] S. R. Garner, et al., "Weka: The waikato environment for knowledge analysis," in Proceedings of the New Zealand computer science research students conference, vol. 1995, 1995, pp. 57–64.

[46] S. Ray, R. H. Turi, "Determination of number of clusters in k-means clustering and application in colour image segmentation," in Proceedings of the 4th international conference on advances in pattern recognition and digital techniques, 1999, pp. 137–143, Citeseer.

[47] T. M. Kodinariya, P. R. Makwana, "Review on determining number of cluster in k-means clustering," *International Journal of Advance Research in Computer Science and Management Studies*, vol. 1, no. 6, pp. 90–95, 2013.

[48] J. C. Dunn, "A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters," Journal of Cybernetics, vol. 3, no. 3, pp. 32–57, 1973, doi: 10.1080/01969727308546046.

[49] W. Zhu, N. Zeng, N. Wang, et al., "Sensitivity, specificity, accuracy, associated confidence interval and roc analysis with practical sas implementations," NESUG proceedings: health care and life sciences, Baltimore, Maryland, vol. 19, p. 67, 2010.

[50] F. Khozeimeh, R. Alizadehsani, M. Roshanzamir, A. Khosravi, P. Layegh, S. Nahavandi, "An expert system for selecting wart treatment method," Computers in biology and medicine, vol. 81, pp. 167– 175, 2017.

Fatima Saadi

She is currently a PhD candidate at the University of Oran 1 and affiliated researcher in Laboratoire d'Informatique d'Oran, Algeria. Her research interests include data mining, case-based reasoning, information retrieval, medical decision support systems, and machine learning.

Baghdad Atmani

He is currently a Full Professor in Computer Science. His field of interest is artificial intelligence and machine learning. His research is based on knowledge representation, knowledge-based systems, CBR, data mining, expert systems, decision support systems and fuzzy logic.

Fouad Henni

He is a teaching researcher in computer science at Mostaganem University, Algeria. His research interests are in the areas of semantic Web services, artificial intelligence, case-based reasoning, and deep learning, with a particular emphasis on applications in medical diagnosis. He is a member of the Computer Science and new Technologies Lab (CSTL).

Improvement of Academic Analytics Processes Through the Identification of the Main Variables Affecting Early Dropout of First-Year Students in Technical Degrees. A Case Study

A. Llauró[1], D. Fonseca[1], E. Villegas[1], M. Aláez[2], S. Romero[2] *

[1] GRETEL, Ramón Llull University, Barcelona (Spain)
[2] Educational Innovation Unit, University of Deusto (Spain)

* Corresponding author: alba.llauro@salle.url.edu (A. Llauró), david.fonseca@salle.url.edu (D. Fonseca), eva.villegas@salle.url.edu (E. Villegas), marian.alaez@deusto.es (M. Aláez), sromeroyesa@deusto.es (S. Romero).

Received 18 February 2022 | Accepted 27 April 2023 | Published 21 June 2023

unir
LA UNIVERSIDAD
EN INTERNET

Abstract

The field of research on the phenomenon of university dropout and the factors that promote it is of the utmost relevance, especially in the current context of the Covid-19 pandemic. Students who have started degrees in the last two years have completed their university studies in periods of lockdown and unlike traditional education, this has often involved taking online classes. In this scenario, the students' motivation and the way they are able to cope with the difficulties of the first year of a university course are very relevant, especially in technical degrees. Previous studies show that a large number of undergraduate students drop out prematurely. In order to act to reduce dropout rates, schools, especially technical schools, should be able to map the entry profile of students and identify the factors that promote early dropout. This paper focuses on identifying, categorizing and evaluating a number of indicators according to the perception of tutors and the field of study, based on the application of quantitative and qualitative techniques. The results support the approach taken, as they show how tutors can identify students at risk of dropping out at the beginning of the course and act proactively to monitor and motivate them.

Keywords

Academic Analytics, Early Dropout, First-Year Students, Learning Analytics, Predictions, Tutoring.

DOI: 10.9781/ijimai.2023.06.002

I. Introduction

Students at almost all levels of education have experienced an abrupt change in their education as a result of the COVID-19 pandemic that began in the spring of 2020 [1]. They have gone from mainly face-to-face education to new online models that are not always well-designed, especially in some technical subjects and in student assessments [2]–[4]. Over the last few academic years, training courses have been constantly modified to be adapted to an online format at specific times. All these changes have had a global impact on the level of education. There are consequences which have already begun to be studied, but that will undoubtedly become evident over the next few years [5]–[7].

School failure is a key factor inherent in educational change. In fact, it is a variable under constant study, and there is a global effort to reduce it [8]. In this respect, authorities have systematically focused their attention on education at pre-university levels. However, recent studies have shown that university dropout has gained importance over the years due to its large increase [9], [10]. The causes of dropout are very diverse, such as economic reasons, family reasons, lack of motivation, etc. [11], [12], but most studies only reflect the data from a descriptive rather than a proactive point of view.

As can be inferred, this fact is more relevant in non-compulsory studies, where students enter voluntarily [13]. The increase in the number of dropouts in the first year of the degree is close to 20% on average (studies put it at 18.72% in 2018 [14]). This increase differs depending on the subject area of the degree, location, etc. [15]–[20]. With the aim of improving this situation, some studies have focused their objectives on the personalisation of student monitoring as a differential factor to reduce the probability of dropout and academic success [21]–[25]. Personalisation is based on the parameterisation of the students' profile and a process of understanding which variables can improve the accompaniment processes [26].

Please cite this article as: A. Llauró, D. Fonseca, E. Villegas, M. Aláez, S. Romero. Improvement of Academic Analytics Processes Through the Identification of the Main Variables Affecting Early Dropout of First-Year Students in Technical Degrees. A Case Study, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 1, pp. 92-103, 2024, http://dx.doi.org/10.9781/ijimai.2023.06.002

This article focuses on a multidisciplinary research project with the aim of parameterising the factors that define the entry profile of undergraduate students at a Spanish national level. The aim is to achieve a better way for tutors to monitor students. This strategy is proposed as an objective to reduce the dropout rate in the first year of study towards a degree. The following research questions have been defined:

RQ1: Is it possible to define an indicator that averages several variables and predicts, in agreement with the tutors, the risk of a first-year undergraduate student dropping out?

RQ2: What level of dropout risk is considered an acceptable range of success for the predictive model?

This document is organized as follows: Section II describes the context in which the study is carried out. Section III explains the methodology followed, as well as the data and variables needed. Section IV shows the data obtained and the associated discussion. Finally, Section V presents the conclusions and possible lines for future research.

II. Context

A. Educational Assessment: Academic Analytics

As we have explained above, the research focuses on identifying and weighing the personal variables that may affect the student's adaptation to the first year of study towards a degree and that may lead to early dropout. Once the identification and weighting process has been completed, relationships will be established between the tutors' perceptions and the results obtained from weighting the variables based on the students' answers, which will allow us to identify possible students at risk and have tutors provide an intervention for the students. In short, we are in front of research that we can circumscribe in the field of evaluation and analysis of educational processes.

While there are no precise definitions in the academic context, training assessment may be defined as "the process of assessing and interpreting organization data collected from university systems for reporting and decision-making purposes" [27], [28]. Learning analytics has arisen as a technique of analysing knowledge acquisition in connection to specific learning objectives, according to this description [29]–[31]. In most situations, it is based on a review of learning outcomes at the conclusion of the training (based on the effectiveness of training, where objectives, content and design of training become the object of evaluation) [32], [33]. "The measurement, gathering, analysis, and reporting of data on learners and their surroundings for the sake of understanding and optimizing learning and the environments in which it happens," according to Ferguson [34].

Academic analytics, in addition to learning analytics, are used to examine the training process at all levels of education, including those that precede training programs and the consequences of these programs [35]. We might suppose that educational data mining [36], [37] is a broad concept that encompasses both learning and academic analytics. Academic analytics is a hybrid method that provides data to higher education institutions to enhance operational and financial decision-making [38]–[40]. While learning analytics is more concerned with course-level and departmental data (to enhance students and professors), academic analytics is more concerned with other factors [41], clearly related to the main topics of our research [42]–[44]:

- learner profiles,
- performance of academics,
- and knowledge flow.

Academic analytics is a field that focuses on the analysis of data from student interactions to improve educational, academic and teaching-related processes [27]. The management of such data provides critical information to educational institutions to make decisions to improve programmes and student tracking and thus maximize student performance [45].

In both research and practice, learning/academic analytics has proven its usefulness in identifying variables that influence learning outcomes and establishing relationships between competencies, educational methodologies and curricular structures [32]. These analyses provide information to personalise courses and to detect at-risk students to provide early intervention. In this way, it is also possible to improve teaching to retain more students throughout the course [46].

B. Tutoring

Tutoring is an activity that has been gaining importance in recent years and has been especially relevant in the period resulting from the COVID-19 pandemic [47], [48], [49]–[51]. The motivation of the student, his or her state of mind derived from the period of lockdown, the need to follow up with students, the difficulty of meeting in-person to complete group work, the review of material needed due to the fact that some subjects are not suited for online learning, and other aspects, are reasons why it's now more important than ever to assist students by providing tutorial services. The pace of work of current students who start a university degree has been weighed down by these last two very educationally complex years, and in a very subjective way, we are noticing a sense of idleness in the daily routine among many of these students. This apathy is caused by a lack of rhythm in their previous studies, constant interruptions and unforeseen changes in the learning model.

Tutorial services are considered a very important intervention in the student's activity throughout their studies. In preuniversity courses, the tutor's main objective is to prevent students from dropping out of school and to identify, in coordination with the teaching staff, learning problems that affect the student [52], [53]. This information is shared with parents and used to initiate the appropriate follow-up.

At the university level, the situation is similar in terms of problem detection and the management of following up with students, but the processes are different [54]. Given that students are adults, the identification of learning problems, their management and the corresponding follow-up are private matters between students and tutors. This means that problems can be more difficult to identify and manage at certain times. Solving the problems of predicting the final marks and combining face-to-face and virtual classes with different student profiles and previous training is a goal for all higher education programs to improve their quality standards [55], [56].

Students recognize the need for generic content in preuniversity studies; however, they do not find a sense of meaning in their choice of university degree , especially when they have chosen a degree with technical-technological-scientific content [57]. This fact, together with the difficulty associated with the educational level, leads to processes of frustration. When other factors are added to this, such as incorrect or insufficiently adapted study habits, the lack of knowledge of how to deal with occasional associated failures, distance from the family environment, greater freedom of movement, etc. [58], the result is that students lack adaptation to university studies.

Therefore, tutoring in the first year of university is of particular importance. The tutor can advise the student on the most critical points of the course, as well as personalise the activity to generate a greater impact and ensure that the student gets through the first year with fewer difficulties [59]–[61]. If the tutor has the ability to collect, analyse and manage data related to the entry profile of his or her students, he or she will be able to anticipate the actions to be taken during the course for those students who may be at risk of dropping out or affected by a situation that may lead to an increase in this risk.

In all these cases, using individual approaches such as coaching, it has been shown that it is possible to address complex situations of the student, starting from the support such as that of a tutor who can help the student discover how to organize him- or herself better, how to take on difficult subjects and how the student perceives the different modes of educational delivery [62]–[64].

C. Methodological Designs

User experience is a discipline that considers the perceptions and responses of people to the behaviour of interaction with a service [65]. It considers both factors linked to the process, and factors related to the emotions of users during the process. The aim is always to achieve good user satisfaction. Therefore, it is based not only on a good process but also on a good experience that encompasses all the points that influence it. Among the possible methodological designs for student monitoring, iterative design stands out as one of the most practical, as it can allow for greater data collection and time management [66].

On the other hand, participatory design actively takes into account all parties involved [67]. Combining iterative and participatory approaches improves the data collection of any study centred on the user who, in the context of our research, is the student. Students are at the centre of the research. The values that are proposed are defined by the students themselves. In this way, it is possible to obtain their profiles and take actions to improve their performance, as well as to help them in the initial adaptation process.

The method applied in the study is based on iterative and participatory design, where the variables selected provide detailed information about the student's profile. From this premise, the User-Centred Design (UCD) [68] methodology is a philosophy that takes into account the user as part of the process of creating the service, providing their motivations, needs or desires during each of the phases [69]. The phases of the iterative process are shown in Fig. 1.

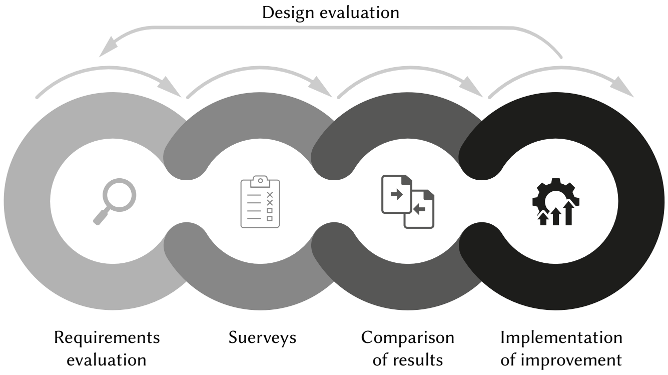


Fig. 1. UCD scheme.

III. Methodology

A. Requirements Assessment

Studies on early dropout may be based on secondary data taken from university computer systems or primary data from a representative sample of students [70]. Those based on secondary data are inexpensive, consider all students and allow multiple analyses by variables. Of this type is the analysis carried out from 1992 to 2006 on 75,830 students pursuing 27 degrees at the University of Granada, which identifies the age of beginning study, the parents' academic degrees and the previous academic results as variables generally associated with early dropout. They also conclude that the profile of the student who drops out is different according to the area of knowledge[71].

More recently, we found data from 2018 in a study where 1071 students from the National Polytechnic School (Quito, Ecuador) were evaluated, which also takes into account the results of university entrance exams. This study concludes that previous academic performance, emotional factors such as attention to emotions and self-esteem are factors that are associated with early dropout [72]. However, this type of analysis has the limitation of ignoring other contextual variables, such as sociopsychological and educational variables, which can be determinants of academic failure.

The study presented in this paper uses primary sources from different faculties and areas of knowledge. It is the students themselves who have directly offered their data to their tutors, which allows the consideration of variables of psychosocial and educational context that may be determinants in the probability of dropout and, on the other hand, will facilitate the approach of interventions that are more tailored to the needs of the students analysed. The disadvantages of this type of study are that it is more expensive and, as we will see later, it requires the design of an instrument for collecting information and the selection of variables a priori.

An example of this type of approach is the 2010 study carried out at Universidad Siglo XXI in Córdoba, Argentina, which concluded that the most influential variable was academic performance, followed by the student's verbal skills [73]; another example is the one carried out in Colombia in 2016, in the department of nursing studies at the Industrial University of Santander, where the variables most related to dropout were academic (low interest in the subject, regular communication with the faculty) and individual types of variables (anxiety, depression and low socioeconomic status) [19].

Finally, it should be noted that there is also room for mixed studies, combining secondary and primary sources. One example is that carried out in Catalan public universities over the course of two academic years (2000 to 2002), which concludes that the first year is the key year that determines the dropout rate and that the most related variables have to do with lack of motivation due to the low quality of the university experience, work or family responsibilities and economic difficulties. A second example of this type is that of the Alfa Guía Project that took place in different European universities over the course of three academic years (2008 to 2011) [72], which, taking into account sociodemographic variables and previous academic performance and vocation, concludes that, among all of the factors, vocation is the most determining. However, this analysis did not incorporate a third block of variables of the area of knowledge or the student's personal adaptation to the university as the most critical aspect.

B. Preliminary Work

As the first phase of UCD, a search, analysis and selection of the different variables to be taken into account to create the student profile was carried out. The variables were selected through a modified Delphi process to determine the content validity of the questionnaire using expert users [74]. The selection of the group of experts to create the first approach was made based on their years of experience in tutoring university students, which in all cases was at least 10 years. There were 12 experts in university tutoring from La Salle URL, 5 tutors from the School of Engineering, 4 tutors from the School of Business, 2 from the School of Architecture and 1 from the School of Digital Arts. In any case, the number of tutors was considered to be sufficient according to the approach used by authors such as Landeta [75] and Cabero and Barroso [76]. Additionally, the time to complete the process, wich was less than two months, is within the limits recommended in the literature [77]–[79]. This method was chosen because its validity in educational research has been widely demonstrated [80], increasing reliability of the study because of the knowledge and consensus of the group consulted [81].

As the first step, a first questionnaire was prepared with 13 items obtained from the review of the existing literature on the research variables and grouped into 3 dimensions: personal data, study habits and motivation. In the first round, the group was asked to rate the items qualitatively. The evaluations were collected personally by e-mail.

Based on the experts' responses, the questionnaire was redesigned to make the study variables measurable, and a second round of validation was subsequently carried out. After the two rounds, the data were statistically processed and returned to the experts to achieve optimal weightings. Based on the answers obtained after the second round and their subsequent analysis, a questionnaire was developed for use in the second phase of UCD.

The 13 data or factors were classified into three main blocks: personal data, study habits and personal motivation.

1. Personal Data

The aim of this block was to collect a set of demographic and social factors that were present before the student began studies toward the degree and that provide information about the student.

- Factor 1, age: this is associated with the variable of the origin of previous studies. There may be students who have begun studies after university entrance exams, who have completed vocational training, who have been in a university program for those over 25 years of age, who have transferred from a different university, who are pursuing a second degree, etc. The combination of this datum together with number 3 (the origin of previous studies) provides information that is related to the motivation variables, as one of the main axes of the instrument [82]–[84].
- Factor 2, gender: Previous research has reported differences in both dropout and academic performance due to this factor [21]. Relationships have been identified between study habits and the gender of the students, as well as a somewhat higher academic performance being associated with study habits in women [85]–[87]. In these studies, it is significant that 25.6% of male students end up dropping out of the selected grade compared to 18.1% of female students [21], [87], which means that gender is an important aspect of the design of the instrument.
- Factor 3, origin of previous studies: as mentioned in the first factor, students with a wide variety of patterns in previous studies can be identified: students from the baccalaureate program, students from the vocational training program, those who have transferred from another university or students over 25 years of age. Previous studies show a higher drop-out rate in those students who have come from a vocational training program or who have completed the entrance exam for students over 25 years of age, while students coming from a baccalaureate program are more likely to change their degrees [13].
- Factor 4, entrance examination marks: depending on the difficulty of the studies, the classic assumption is to associate a lower entrance examination mark with a higher probability of student dropout. The aim of incorporating this factor is to see if there are differences depending on the field of knowledge and to corroborate whether previous data in other fields are confirmed. For example, in the area of health sciences, 89.1% of students obtain an entrance examination mark of 7.5 or higher and have a dropout rate of 11.11%, while in architecture and engineering, the rate of students with an entrance examination mark of 7.5 is only 65.4% and they have a dropout rate of 25.65% [14]. The grade is positioned as a determining factor and, in conjunction with degree changes, it can be highly significant in the parameterisation of the student profile [70], [88].
- Factor 5, country of precedence of compulsory studies: the ORCE (Organisation for Economic Co-operation and Development, which is responsible for analysing the academic performance of each of the member countries) has found that Spain is within the average of the member countries in terms of the diversity of origin of students [89]. In these studies, origin is considered a highly significant factor in the possibility of dropout, in many cases resulting from a lack of integration of the student (both academic and social) and/or due to bureaucratic or economic obstacles that complicate the student's day-to-day life.

2. Study Habits

The section on study habits includes factors related to the student's way of studying: how he or she had worked before entering university and how he or she structured and managed time and resources when studying. In this sense, three types of data have been defined that make up the instrument on work:

- Factor 6: I do homework at the last minute.
- Factor 7: I schedule myself daily study time.
- Factor 8: how many days I study before exams or final papers.

The need for these data is referenced in previous studies where a correlation has been observed between academic performance and study habits [90]. It was identified that up to 50% of students do not have good study habits and do not follow the course of study in a subject. For this reason, they need to undertake a process of adaptation during the first year of the degree to enable them to pass the course [91], taking into account the higher level of difficulty of the subjects compared to those in their previous years of study.

3. Motivation

The third and last block of data was concerned with the concept of student motivation. The aim is to evaluate the attitude of students towards continuing their studies despite the difficulties they may encounter during a course:

- Factor 9, students' conviction of their choice of degree: This question is a very relevant aspect of this study that is complemented by the lack of information on the degree in which the students have enrolled. Previous studies have identified that approximately 45% of the students who start a degree do not have the necessary information about it, and 24.3% have not selected it as their first choice [92]. Vocation is one of the most important aspects, together with conviction (related to information about the studies) of the student in regard to making the right choice of degree programme. Students who do not have a strong vocation for their chosen field of study are twice as likely to drop out [93].
- Factor 10, first choice: the degree selected as first choice demonstrates that students have a feeling of conviction in regard to studying. Previous work has identified that a large proportion of students who dropped out were in a field of study that was not their first choice. Up to 82% of students who dropped out of their chosen degree program did not select it as the first choice, with 49% of these students dropping out during the first year [94].
- Factor 11, branch of education: The Conference of Rectors of Spanish Universities (CRUE) has studied dropout rates according to the corresponding subject area. Some examples are that 25.63% of architecture and engineering students drop out in the first year, 22.57% of students in the humanities and arts, 16.81% of students in the social or legal area and 11.11% of students in health sciences [14]. Studying these data in a limited context (studies at La Salle, Ramon Llull University) in four different degree courses in different subject areas would make it possible to corroborate, qualify and extend previous studies, generating an innovative contribution to the field of research on evaluation methods in didactics.
- Factor 12, distance to the university: many of the students who come to the university come from faraway places. Due to this

distance, these students may stay in a university residence, in a shared flat, in a flat on their own, with relatives, or make very diverse journeys every day. Journeys of one up to two or more hours to get to the centre are factors that can directly influence students' academic performance and motivation [95].

- Factor 13, scholarship: In the case of public schools and to even a greater extent in private schools (such as the one in this study), obtaining and maintaining a scholarship is essential to support the cost of university studies. The stress that can be caused by maintaining an average grade necessary to fulfil the requirements of a scholarship for excellence, or the need to work to earn money, can influence the student's performance, cause greater fatigue and afffect motivation. This factor is therefore incorporated into the instrument. As seen from previous studies, 18.6% of nonscholarship students eventually drop out of their chosen degree, while 14.2% of scholarship students drop out. In private universities, differences are also observed in the dropout rates of students with and without scholarships, with a dropout rate of 13.2% of students without scholarships and 10.1% of students with scholarships [96].

C. Questionnaire

An instrument was created in the form of a questionnaire to characterise the student profile based on the significant data identified in the previous process. On the one hand, students were asked to give a first approximation of their personal characteristics (quantitative approximation considering the number of samples). On the other hand, first-year tutors were asked to give a qualitative weight for each data or factor according to its level of importance. Weights were given by the tutors to each of the 13 factors studied, and the aim was for the weight to be applied to the students' assessment of each factor in their personal situation. This way, students classify themselves, and their risk of dropping out can be identified.

1. Students

The aim is to obtain the data from new students at the first stage in the university: the "Welcome Week" (see Table I). In this initial week of the course, students are welcomed by their assigned tutors, who introduce them to the physical spaces, digital systems and management and monitoring aspects of the course. This creates a link that allows for greater empathy between students and tutors in terms of monitoring and action.

TABLE I. Student Survey

#	Question
Data 1	Date of birth
Data 2	Gender
Data 3	What did you study?
Data 4	Average mark for selectivity or other studies (Example: 8.85)
Data 5	Where did you study in high school or the last compulsory course? (Country)
Data 6	Do you do the homework at the last minute or when they are sent to you?
Data 7	Do you study and review the subject every day?
Data 8	How many days before an exam do you start studying?
Data 9	How sure are you of the degree you have chosen?
Data 10	Was studying for this career your first option?
Data 11	What field does your career belong to?
Data 12	How long does it take you to get to the university?
Data 13	Do you have a scholarship?

2. Tutors

As mentioned above, a questionnaire was carried out among the tutors with the aim of finding the weighting method for each factor/data component of the predictive instrument created. A total of 11 first-year tutors from the four fields of study, which were engineering, architecture, business and digital arts, participated. In this survey, the tutors ranked the 13 pieces of information required from the students from most to least important. Fig. 2 shows the differences obtained in the weighting of the importance of each factor by subject area.

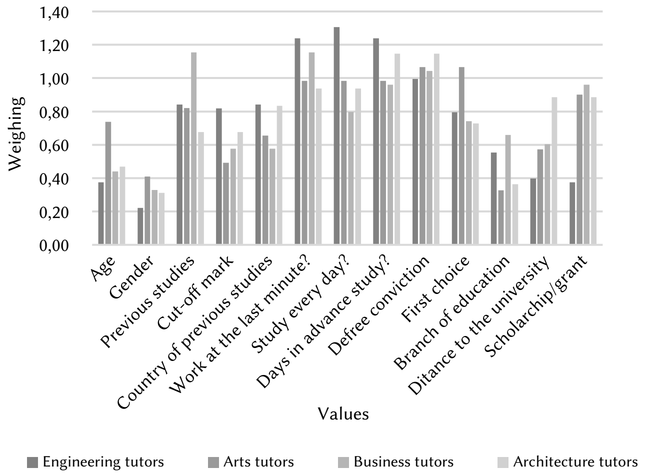


Fig. 2. Weighting by areas of knowledge.

The average for each factor, obtained from the responses of the 11 tutors without distinction of the area of knowledge, was used for this first iteration of the research. The final weighting for each factor is shown in Table II.

TABLE II. Final Weights and Standard Deviation

Value	Analytical technique	
	μ	σ
Age	0.45	0.26
Gender	0.29	0.17
Previous studies	0.91	0.32
Cut-off mark	0.68	0.33
Country of previous studies	0.73	0.26
Work at the last minute?	1.13	0.25
Study every day?	1.04	0.30
Days in advance study?	1.11	0.24
Degree conviction	1.04	0.24
First choice	0.80	0.36
Branch of education	0.53	0.37
Distance to the university	0.57	0.31
Scholarship/grant	0.71	0.38
	Base MVA=10	

D. Results Comparison

Once the tutors had assessed the weight of each piece of information and each of them had been weighted to obtain the prediction instrument, the students' responses were integrated. To do this, the next step was the classification of the students by all their first-year tutors.

The aim was to establish categories according to the drop-out risk perceived by the tutors after the first six weeks of class and before the mid-semester checkpoint.

Furthermore, the aim is that this prediction made independently by the tutor is repeated at the end of the first semester and at the second semester checkpoint so that a relationship can be established between the evolution perceived by the tutor without marks, at mid-term, and facing the final stretch, with the perception of the student in his or her initial state.

The classification made by the tutors at the three points in time described is based on a traffic light with three levels according to the low, medium or high risk of the student dropping out. These perceptions are then compared with the result of the initial perception after the survey of each student.

E. Implementation of Improvements

The last phase of the iterative process is based on adjusting the weights of each data item according to the results obtained from the comparison between the data collected from the student and the perception of the tutors. In this way, a more approximate estimation can be achieved, and the performance of the instrument can be improved.

The aim is not only to improve the weighting of each piece of data but also to provide the tutor with an active monitoring and intervention tool for students at risk of dropping out. Moreover, as an iterative methodology, over the following academic years, the aim is to integrate new students, tutors and grades so that an increasingly fine-tuned instrument can be developed that can be used with modifications of weights depending on the field of study.

Fig. II shows that there are differences between the perceptions of each factor in terms of its relevance to the risk of dropout according to subject area.

IV. Case Study

A. The Sample

The results obtained come from a sample of 309 new undergraduate students from the four previously identified areas of knowledge of La Salle - Ramon Llull University. The average age of the sample was 18.96 years old. 36% of the students surveyed were female, 63% were male, and 1% preferred not to specify their gender.

TABLE III. Personal Data Collected

Value	Answer	Analytic	
		n	%
Age	<20	246	80%
	>=20	58	19%
	>=25	5	2%
Gender	Female	110	36%
	Male	196	63%
	Unspecified	3	1%
Previous studies	Secondary school	246	80%
	Higher Vocational Training	34	11%
	University transfer	25	8%
	Other	4	1%
Cut-off mark	>= 7.5	204	66%
	>= 6	81	26%
	>= 5	19	6%
	<5	5	2%
Country of previous studies	>Spain	254	82%
	<Spain	7	2%
	<<Spain	48	16%

80% of the students came from a baccalaureate program, compared to 11% who come from vocational training and 8% who were transfer students from another university. 55% percent of the respondents were from the ICT Engineering and Technology area, 16% were from Business and Management, 16% were from Digital Arts, Animation and VFX and 13% were from Architecture.

74% of the total sample of students considered the selected degree to be their first choice, and the average score for the level of security of the selected degree was 4.28 out of 5.

Finally, it should be noted that 53% of the students had a financial grant, which represented an increase in funding for grants due to the pandemic.

Table III shows the results obtained in the personal data area, showing the total number of students who chose each answer and the corresponding percentage.

Table IV Shows the Results Obtained in the Area of Study Habits. It can be Observed how the Different Students Work.

TABLE IV. Study Habits Data Collected

Value	Answer	Analytic	
		n	%
Work at the last minute?	No	208	67%
	Yes	101	33%
Study every day?	No	198	64%
	Yes	111	36%
Number of days in advance to study?	More than a week	29	9%
	One week	80	26%
	3 to 5 days	126	41%
	1 to 2 days	69	22%
	The day before	5	2%

Table V shows the results obtained in the area of student motivation, the confidence with which the student chose the degree and the other factors identified.

TABLE V. Motivational Data Collected

Value	Answer	Analytic	
		n	%
Degree conviction	5	127	41%
	4	150	49%
	3	24	8%
	2	6	2%
	1	2	1%
First choice	No	81	26%
	Yes	228	74%
Branch of education	Business	48	16%
	Art	49	16%
	Architecture	41	13%
	Engineering	171	55%
Distance to the university	Less than 15 min	62	20%
	15 to 30 min	57	18%
	30 to 45 min	57	18%
	45 to 60 min	66	21%
	More than 1 h	67	22%
Scholarship/grant	No	144	47%
	Yes	165	53%

TABLE VI. Example of Users With Different Weightings

Name	**User 1**	**User 51**	**User 67**	**User 184**	**User 203**
Age	19	18	18	30	18
Gender	Male	Male	Female	Male	Female
Previous studies	Baccalaureate	Baccalaureate	Baccalaureate	Other	Baccalaureate
Cut-off Mark	7.33	9.2	6.8	7	7.5
Country	Spain	Spain	Spain	Argentine	Spain
Last minute?	No	No	No	No	Yes
Every day?	Yes	Yes	No	Yes	No
Days in advance to study	One week	One week	One week	More than one week	3 to 5 days
Conviction	5	5	4	4	3
First option	No	Yes	Yes	No	No
Degree	Business	Architecture	Engineering	Engineering	Arts
Time-distance	30-45 min	Less than 15 min	More than 1 h	30-45 min	45-60 min
Grant	Yes	Yes	No	Yes	No
Score	**8.722**	**7.856**	6.386	6.109	4.597

B. Weighting of Results

The data obtained, subsequently weighted, are ranked according to the final mark obtained. Each data point receives either the full weighting or a part of it, depending on the weight of the data analysed in previous research. For example, in the case of the conviction of the selected degree, 5 receives the full weighting and 1 receives nothing; for other values, the proportional part is assigned. Responses with two options are either fully weighted or not weighted at all. Finally, all the weights obtained for each of the values are added together to obtain a single value for ranking. Those students with a score below 5 are considered to be at very high risk of dropping out, between 5 and 6, high risk, between 6 and 7 medium risk, and finally those with a score above 7 are considered to be at low or very low risk of dropping out. Table VI shows the classification of the different profiles. Each of the users is part of a sample response according to their risk. User 1 would have a very low probability of dropout and user 203 would have a very high probability of dropout. The value of the score is the result of weighting the different responses.

C. Results Analysis

Based on the tutors' initial classification of the risk of their students dropping out (low, medium, high), it was observed that after a few weeks, 1% of the students dropped out of the degree course. The explanation for this fact is based on possible double enrolments and the inclusion of students who had been accepted at a university but finally decided not to attend. Then, 12% of the students were identified as being at high risk of dropping out, 33% had a medium risk, and 54% had no apparent risk. In monitoring processes with tutors, it was found that they carried out monitoring actions in the first instance, with 12% of students identified as being at high risk of dropping out through immediate and regular meetings. The percentage distribution of students by area and perceived drop-out risk is shown in Table VII. Overall, 16.83% of students are at a very high risk of dropping out of the degree, 20.79% are at high risk, and 26.07% are at medium risk. On the other hand, 26.4% have a low risk, and 9.9% have a very low risk of dropping out.

TABLE VII. Weighting of the Initial Surveys of Students in the Different Areas

	ARTS	BUSIN	ENG	ARQ	TOTAL
Very low	16.33%	25.00%	4.14%	8.11%	9.90%
Low	32.65%	33.33%	23.08%	24.32%	26.40%
Medium	24.49%	18.75%	26.04%	37.84%	26.07%
High	12.24%	18.75%	25.44%	13.51%	20.79%
Very high	14.29%	4.17%	21.30%	16.22%	16.83%

D. Results Obtained

The results obtained are compared with the perception of the tutors at the beginning and at the end of the course (see Table VIII). At the end of the course, a new study is carried out to show student enrolment in the new year of the course. A comparison is made between the end of the course and the initial questionnaire in order to know the exact results obtained by those students who started their studies. The similarity of the results drops to 53% due to the different actions carried out by the tutors during the course to mitigate the drop-out rate.

TABLE VIII. Comparison Between Survey and Perception of Tutors

	Beginning of the course	End of the course
Same	72%	53%
Medium	19%	24%
Opposite	9%	23%

Finally, Table IX shows that a total of 32.79% of the students initially classified with a very high risk of dropping out finally left the grade selected. Similarly, 18.03% of the students who were classified as high risk dropped out. On the other hand, only 3.28% of students classified as very low risk dropped out.

TABLE IX. Initial Prediction of Students Who Finally Dropped Out

	ARTS	BUSIN	ENG	ARQ	Total
Very low	20.00%	0.00%	0.00%	20.00%	3.28%
Low	60.00%	0.00%	16.33%	20.00%	19.67%
Medium	0.00%	100.00%	26.53%	20.00%	26.23%
High	0.00%	0.00%	20.41%	20.00%	18.03%
Very high	20.00%	0.00%	36.73%	20.00%	32.79%

Once the final results of the students had been obtained, those who finally dropped out were selected to analyse the profile and observe which of the variables are the most important in predicting early dropout. Table X shows the personal data of those students who finally dropped out, as well as the percentage they represent with respect to the total number of students who chose the same answer.

TABLE X. Personal Data of Students Who Dropped Out

Value	Answer	Analytic		
		n.total	n.dropout	%
Age	<20	246	46	18.70%
	>=20	58	10	17.24%
	>=25	5	4	80.00%
Gender	Female	110	14	12.73%
	Male	196	46	23.47%
	Unspecified	3	0	0.00%
Previous studies	Secondary school	246	47	19.11%
	Higher Vocational Training	34	6	17.65%
	University transfer	25	4	16.00%
	Other	4	3	75.00%
Cut-off mark	>= 7.5	204	36	17.65%
	>= 6	81	17	20.99%
	>= 5	19	4	21.05%
	<5	5	3	60.00%
Country of previous studies	>=Spain	254	47	18.50%
	<Spain	7	4	57.14%
	<<Spain	48	9	18.75%

From the results obtained and summarized in Table X, it can be observed that 80% of the students over 25 years of age dropped out of the course, and the dropout rate among men was practically double that of women, with 75% of dropouts being identified in those students who came from other access modalities than those indicated, although given the number of these exceptional cases, we cannot consider it to be significant. On the other hand, students with a grade lower than 5 accounted for 60% of dropouts, followed by those with a grade between 5 and 6 (21.05%) and those with a grade between 6 and 7 (20.99%). Finally, another relevant fact is that students from abroad (not Spain) account for 57.14% of dropouts, which may suggest problems of rootedness, homesickness or lack of adaptation to a higher academic level. Table XI analyses the habits of the students:

TABLE XI. Data on the Study Habits of Dropouts

Value	Answer	Analytic		
		n.total	n.dropout	%
Work at the last minute?	No	208	34	16.35%
	Yes	101	26	25.74%
Study every day?	No	198	34	17.17%
	Yes	111	26	23.42%
Days in advance study?	More than a week	29	4	13.79%
	One week	80	14	17.50%
	3 to 5 days	126	23	18.25%
	1 to 2 days	69	17	24.64%
	The day before	5	2	40%

Analysing the results of the study habits reported by the students, it is observed that among the confirmed dropouts, there was a tendency to work at the last minute, which was almost double that of those who plan ahead of time, with approximately 25% of the dropouts being students who report such habits. Finally, Table XII shows the results about the third section of the questionnaire, related to the motivation of the student:

TABLE XII. Motivational Data of Dropouts

Value	Answer	Analytic		
		n.total	n.dropout	%
Degree conviction	5	127	21	16.54%
	4	150	29	19.33%
	3	24	4	16.67%
	2	6	4	66.67%
	1	2	2	100.00%
First choice	No	81	25	30.86%
	Yes	228	35	15.35%
Branch of education	Business	48	2	4.17%
	Art	49	5	10.20%
	Architecture	41	5	12.20%
	Engineering	171	48	28.07%
Distance to the university	Less than 15 min	62	11	17.74%
	15 to 30 min	57	9	15.79%
	30 to 45 min	57	12	21.05%
	45 to 60 min	66	16	24.24%
	More than 1 h	67	12	17.91%
Scholarship/grant	No	144	36	25.00%
	Yes	165	24	14.55%

When analysing the motivation of the students who dropped out (Table XII), it can be seen that 100% of the students who had marked 1 (low) were convinced of their choice of degree dropped out, followed by 66.67% of the students who had marked 2. A total of 30.86% of the students who had said that the degree was not their first choice finally dropped out. It can also be observed that 28.07% of engineering students dropped out, followed by architecture students with 12.20%. With regard to the distance to the university, there were few differences, but 24.24% of the students who dropped out were between 45 and 60 minutes away from the university, followed by those who were between 30 and 45 minutes away, with 21.05%. Finally, 25% of students who did not have a grant left the degree program in the first year. More specific profiles can be created depending on the area, but it should be noted that, of the total number of dropouts, most were in engineering, and 100% of the dropouts in architecture were women.

As shown in the first phase of the method, it is possible to detect and define different indicators that, when averaged, give a dropout risk value to each of the new students, thus resolving the RQ1 formulated in the research in the affirmative. Table IX shows the evolution of the comparison during the course, where in the middle of the course, the perception of the tutors coincides by 72% in relation to the average of the different indicators.

Finally, if we analyse only the dropouts at the end of the course and according to the result obtained in the initial weighting, we can conclude that those students identified with an average below 5 are potential dropouts. In detail, Table IX provides us with the complementary information that those students with an average between 5 and 7 have a very high risk of dropping out; therefore, urgent actions are necessary at the tutorial level. Above 7, the risk of the student dropping out during the first-year decreases, and this value is the limit to be monitored. This analysis confirms the feasibility of the second proposed RQ2 and establishes a given mean that identifies the student's dropout risk.

V. Conclusion and Future Research Lines

This project shows that the initial survey can be used to create the student's profile, as well as to detect the variables necessary to predict possible dropout. Thanks to this survey, the tutor reduces the initial effort and helps the student to anticipate his or her work. Initially, the tutor does not receive information about the student until they meet. By means of the survey, the tutor can receive the data at the beginning of the course so that he or she can speed up his or her work and detect those students at risk as early as possible.

The study has proven the high similarity between the initial prediction extracted from the student questionnaire and the perception at the beginning of the course of each of the tutors about their students. This similarity will be corroborated with the iteration of the study to verify the degree of correlation in future iterations, but it has already allowed us to affirm that it is possible to define an indicator that by averaging various factors, we can use to predict the risk of dropout in relation to the weighting of these factors by the tutors. It has also been proven that this value decreases throughout the year, which is presumably due to the work carried out by the tutors from the initial state to reduce the risk of abandonment.

It can be observed that those students classified as having a very low probability of dropping out do not have a high degree of dropout, unlike those who were initially classified as having a high or very high degree of dropout. It could be said that the weighting limit would be at a value of 7, and above this value, they would be at a medium risk of dropping out. Those cases that are in the middle should be taken into account; they may drop out for reasons that were not detected at the beginning and can only be mitigated with tutorial actions.

These aspects are fundamental in the university environment, as it has been demonstrated that there is a large increase in dropouts at the beginning of the degree program, even in the first months. As this process iterates and is evaluated over the years, possible dropouts due to frustration, lack of motivation or lack of knowledge about the selected degree can be controlled.

Evaluating the results of the demographic aspects, the study demonstrated that age and university entrance exam score, as well as motivation in the selection of the degree, are aspects that are highly influential on student dropout. Regarding study habits, it is confirmed that students with less regular study habits and/or those who study at the last minute are also identified by the instrument as being at high risk.

With the information provided by the instrument, the tutor has tools for follow-up and action to help the at-risk student. In this sense, once those students who stand out negatively either totally or partially by looking at indicated factors are identified, the following actions are proposed by the tutor: a) imminent follow-up meeting, even before the first checkpoint, b) creation of study guidelines, c) approaching the student to invite him or her study and/or reinforcement groups, d) establish contact with the teachers for a closer follow-up of the student, and e) proposal of a personalized follow-up based on the application of coaching techniques. All these actions have been demonstrated to be effective in past courses and in pilot activities that reduce the dropout rate of students at risk [63], since there is a rapprochement between the student, teachers and tutors that minimizes negative emotional aspects such as embarrassment, shyness, feelings of frustration, and fear and improves the motivation, tranquillity and satisfaction of the student. Getting the student to overcome the invisible barrier that separates him or her from teachers and subjects that he or she does not handle or manage well is the first step for the effective reduction of early dropout.

Another aspect to take into account to reduce dropout rates is educational reform, considering the development of emotional competencies for the better development of students. These competencies help university students face challenges more easily, thus promoting entrepreneurship and reducing dropout rates. [85], [97]–[99].

For the next iterations, a survey will be generated where more specific data provided by the student will be needed. These data will be selected by the tutors of the different areas and weighted by tutors themselves according the field of knowledge. In this way, tutors will use their collective experience to create exhaustive profiles of each of the new students according their personal data and their relation with the educational field.

The student's personal and educational aspects as well as his or her motivation and study habits will be taken into account. The aim is to create a report that is detailed and has as much information as possible to be able to weigh the data and obtain the most relevant information from each of them.

References

[1] F. J. García-Peñalvo, "Digital Transformation in the Universities: Implications of the COVID-19 Pandemic," Transformación digital en las universidades: Implicaciones de la pandemia de la COVID-19, Feb. 2021, Accessed: Mar. 02, 2022. [Online]. Available: https://repositorio.grial.eu/handle/grial/2230

[2] T. Knopik and U. Oszwa, "E-cooperative problem solving as a strategy for learning mathematics during the COVID-19 pandemic," Education in the Knowledge Society (EKS), vol. 22, pp. e25176–e25176, Dec. 2021, doi: 10.14201/eks.25176.

[3] F. J. García-Peñalvo, A. Corell, V. Abella-García, and M. Grande, "Online assessment in higher education in the time of COVID-19," Education in the Knowledge Society, vol. 21, 2020.

[4] F. J. García-Peñalvo, A. Corell, V. Abella-García, and M. Grande-de-Prado, "Recommendations for Mandatory Online Assessment in Higher Education During the COVID-19 Pandemic," in Radical Solutions for Education in a Crisis Context: COVID-19 as an Opportunity for Global Learning, D. Burgos, A. Tlili, and A. Tabacco, Eds., in Lecture Notes in Educational Technology. Singapore: Springer, 2021, pp. 85–98. doi: 10.1007/978-981-15-7869-4_6.

[5] J. P. Azevedo, A. Hasan, D. Goldemberg, S. A. Iqbal, and K. Geven, "Simulating the Potential Impacts of COVID-19 School Closures on Schooling and Learning Outcomes: A Set of Global Estimates," World Bank, Washington, DC, Working Paper, Jun. 2020. doi: 10.1596/1813-9450-9284.

[6] J. G. Fuenmayor and C. M. Bolaños, "Estrategias de aprendizaje para mitigar la deserción estudiantil en el marco de la COVID-19," SUMMA. Revista disciplinaria en ciencias económicas y sociales, vol. 2, pp. 49–55, Sep. 2020, doi: 10.47666/summa.2.esp.06.

[7] G. Jacobo-Galicia, A. I. Máynez-Guaderrama, and J. Cavazos-Arroyo, "Miedo al Covid, agotamiento y cinismo: su efecto en la intención de abandono universitario," European Journal of Education and Psychology, vol. 14, no. 1, pp. 1–18, Mar. 2021, doi: 10.32457/ejep.v14i1.1432.

[8] G. W. Dekker, M. Pechenizkiy, and J. M. Vleeshouwers, "Predicting Students Drop Out: A Case Study.," International Working Group on Educational Data Mining, 2009.

[9] B. Pérez, C. Castellanos, and D. Correal, "Predicting student drop-out rates using data mining techniques: A case study," in IEEE Colombian Conference on Applications in Computational Intelligence, Springer, 2018, pp. 111–125.

[10] D. Bustamante and O. Garcia-Bedoya, "Predictive Academic Performance Model to Support, Prevent and Decrease the University Dropout Rate," in International Conference on Applied Informatics, Springer, 2021, pp. 222–236.

[11] Á. Choi de Mendizábal and J. Calero Martínez, "Determinantes del riesgo de fracaso escolar en España en PISA-2009 y propuestas de reforma," Revista de Educación, no. 362, 2013, doi: 10.4438/1988-592X-RE-2013-362-242.

[12] J. M. Guio-Jaimes and A. Choi-de-Mendizábal, "The Evolution of School Failure Risk During the 2000 Decade in Spain: Analysis of PISA Results

with a Two-Level Logistic Model," Evolución del riesgo de fracaso escolar en España durante la década del 2000: Análisis de los resultados de PISA con un modelo logístico de dos niveles, 2014, Accessed: Mar. 01, 2022. [Online]. Available: https://dadun.unav.edu/handle/10171/36784

[13] E. Corominas Rovira, "La transición de los estudios universitarios: Abandono o cambio en el primer año de Universidad," Revista de investigación educativa, RIE, vol. 19, no. 1. pp. 127–152, 2001.

[14] J. A. Pérez García, J. Hernández Armenteros, and Conferencia de Rectores de las Universidades Españolas, La universidad española en cifras 2017/2018. Madrid: CRUE, 2020.

[15] C. M. Fourie, "Risk factors associated with first-year students' intention to drop out from a university in South Africa," Journal of Further and Higher Education, vol. 44, no. 2, pp. 201–215, 2020.

[16] S. C. Wolter, A. Diem, and D. Messer, "Drop-outs from S wiss Universities: an empirical analysis of data on all students between 1975 and 2008," European Journal of Education, vol. 49, no. 4, pp. 471–483, 2014.

[17] M. S. A. Taipe and D. M. Sánchez, "Prediction of university dropout through technological factors: a case study in Ecuador," p. 7.

[18] J. A. Z. Araya and F. J. V. Madrigal, "Factors associated with dropping out of the program for Bachelor's and Licentiate's Degrees in Mathematics Teaching at the Universidad Nacional de Costa Rica (UNA): Evidence from the 2016 Student Cohort," Uniciencia, vol. 32, no. 2 (July-December), pp. 111–126, 2018.

[19] C. V. Porras, D. I. Parra, and Z. M. R. Díaz, "Factores relacionados con la intención de desertar en estudiantes de enfermería.: Factors relating to nurse students intending to drop out.," Revista Ciencia y Cuidado, pp. 86–97, Jan. 2019, doi: 10.22463/17949831.1545.

[20] J. Gairín, X. M. Triado, M. Feixas, P. Figuera, P. Aparicio-Chueca, and M. Torrado, "Student dropout rates in Catalan universities: profile and motives for disengagement," Quality in Higher Education, vol. 20, no. 2, pp. 165–182, May 2014, doi: 10.1080/13538322.2014.925230.

[21] Á. Choi and J. Calero, "Early School Dropout in Spain: Evolution During the Great Recession," in European Youth Labour Markets, M. Á. Malo and A. Moreno Mínguez, Eds., Cham: Springer International Publishing, 2018, pp. 143–156. doi: 10.1007/978-3-319-68222-8_10.

[22] J. D. Corral, J. L. González-Quejigo, and M. Villasalero, "Análisis del abandono universitario en la universidad de castilla-la mancha: resultados del proyecto Alfa Guía," Congresos CLABES, 2015, Accessed: Mar. 01, 2022. [Online]. Available: https://revistas.utp.ac.pa/index.php/clabes/article/view/1097

[23] A. G. de Fanelli and C. A. de Deane, "Abandono de los estudios universitarios: dimensión, factores asociados y desafíos para la política pública," Revista Fuentes, no. 16, Art. no. 16, 2015.

[24] C. Burgos, M. L. Campanario, D. de la Peña, J. A. Lara, D. Lizcano, and M. A. Martínez, "Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout," Computers & Electrical Engineering, vol. 66, pp. 541–556, 2018, doi: https://doi.org/10.1016/j.compeleceng.2017.03.005.

[25] Méndez-Ortega, L. A. Urbina-Nájera, A. B., "Predictive Model for Taking Decision to Prevent University Dropout," International Journal Of Interactive Multimedia And Artificial Intelligence, vol. In press, no. In press, pp. 1–9, 2022, doi: http://doi.org/10.9781/ijimai.2022.01.006.

[26] Alonso-Misol Gerlache, H., Moreno-Ger, P., & de-la-Fuente Valentín, L., "Towards the Grade's Prediction. A Study of Different Machine Learning Approaches to Predict Grades from Student Interaction Data," International Journal Of Interactive Multimedia And Artificial Intelligence, vol. In press, no. In press, pp. 1–9, 2022, doi: http://doi.org/10.9781/ijimai.2021.11.007.

[27] D. A. Filvà, M. A. Forment, F. J. García-Peñalvo, D. F. Escudero, and M. J. Casañ, "Clickstream for learning analytics to assess students' behavior with Scratch," Future Generation Computer Systems, vol. 93, pp. 673–686, Apr. 2019, doi: 10.1016/j.future.2018.10.057.

[28] F. J. García-Peñalvo, "Avoiding the Dark Side of Digital Transformation in Teaching. An Institutional Reference Framework for eLearning in Higher Education," Sustainability, vol. 13, no. 4, Art. no. 4, Jan. 2021, doi: 10.3390/su13042023.

[29] R. Ferguson, "Learning analytics: drivers, developments and challenges," International Journal of Technology Enhanced Learning, vol. 4, no. 5–6, pp. 304–317, Jan. 2012, doi: 10.1504/IJTEL.2012.051816.

[30] A. H. Duin and J. Tham, "The Current State of Analytics: Implications for Learning Management System (LMS) Use in Writing Pedagogy," Computers and Composition, vol. 55, p. 102544, Mar. 2020, doi: 10.1016/j.compcom.2020.102544.

[31] A. Y. Q. Huang, O. H. T. Lu, J. C. H. Huang, C. J. Yin, and S. J. H. Yang, "Predicting students' academic performance by using educational big data and learning analytics: evaluation of classification methods and learning logs," Interactive Learning Environments, vol. 28, no. 2, pp. 206–230, Feb. 2020, doi: 10.1080/10494820.2019.1636086.

[32] R. S. Baker and P. S. Inventado, "Educational Data Mining and Learning Analytics," in Learning Analytics: From Research to Practice, J. A. Larusson and B. White, Eds., New York, NY: Springer, 2014, pp. 61–75. doi: 10.1007/978-1-4614-3305-7_4.

[33] A. Balderas, M. Palomo-Duarte, J. Antonio Caballero-Hernández, M. Rodriguez-Garcia, and J. Manuel Dodero, "Learning Analytics to Detect Evidence of Fraudulent Behaviour in Online Examinations.," International Journal of Interactive Multimedia & Artificial Intelligence, vol. 7, no. 2, 2021.

[34] R. Ferguson, "Learning analytics: Drivers, developments and challenges," International Journal of Technology Enhanced Learning, vol. 4, no. 5–6, pp. 304–317, 2012, doi: 10.1504/IJTEL.2012.051816.

[35] G. Siemens and P. Long, "Penetrating the Fog: Analytics in Learning and Education," EDUCAUSE Review, vol. 46, no. 5, p. 30, 2011.

[36] M. Á. Conde and Á. Hernández-García, "A promised land for educational decision-making? present and future of learning analytics," in Proceedings of the First International Conference on Technological Ecosystem for Enhancing Multiculturality, in TEEM '13. New York, NY, USA: Association for Computing Machinery, Nov. 2013, pp. 239–243. doi: 10.1145/2536536.2536573.

[37] H. Waheed, S.-U. Hassan, N. R. Aljohani, J. Hardman, S. Alelyani, and R. Nawaz, "Predicting academic performance of students from VLE big data using deep learning models," Computers in Human Behavior, vol. 104, p. 106189, Mar. 2020, doi: 10.1016/j.chb.2019.106189.

[38] P. J. Goldstein and R. N. Katz, Academic analytics: The uses of management information and technology in higher education, vol. 8. Educause, 2005.

[39] P. Baepler and C. Murdoch, "Academic Analytics and Data Mining in Higher Education," International Journal for the Scholarship of Teaching and Learning, 2010, doi: 10.20429/ijsotl.2010.040217.

[40] J. P. Campbell, P. B. DeBlois, and D. G. Oblinger, "Academic Analytics: A New Tool for a New Era," Educause Review, 2007.

[41] M. A. Chatti, A. L. Dyckhoff, U. Schroeder, and H. Thüs, "A reference model for learning analytics," International Journal of Technology Enhanced Learning, vol. 4, no. 5–6, pp. 318–331, Jan. 2012, doi: 10.1504/IJTEL.2012.051815.

[42] E. P. Camarilla, D. F. Escudero, and N. M. Audí, "Relationship between specific professional competences and learning activities of the building and construction engineering degree final project," The International journal of engineering education, vol. 34, no. 3, pp. 924–939, 2018.

[43] E. Peña, D. Fonseca, and N. Martí, "Relationship between learning indicators in the development and result of the building engineering degree final project," in ACM International Conference Proceeding Series, 2016, pp. 335–340. doi: 10.1145/3012430.3012537.

[44] D. Fonseca, N. Martí, E. Redondo, I. Navarro, and A. Sánchez, "Relationship between student profile, tool use, participation, and academic performance with the use of Augmented Reality technology for visualized architecture models," Computers in Human Behavior, vol. 31, no. 1, pp. 434–445, 2014, doi: 10.1016/j.chb.2013.03.006.

[45] U. bin Mat, N. Buniyamin, P. M. Arsad, and R. Kassim, "An overview of using academic analytics to predict and improve students' achievement: A proposed proactive intelligent intervention," in 2013 IEEE 5th Conference on Engineering Education (ICEED), Dec. 2013, pp. 126–130. doi: 10.1109/ICEED.2013.6908316.

[46] S. Palmer, "Modelling engineering student academic performance using academic analytics," International journal of engineering education, vol. 29, no. 1, pp. 132–138, Jan. 2013.

[47] N. Yusuf, "The Effect of Online Tutoring Applications on Student Learning Outcomes during the COVID-19 Pandemic," ITALIENISCH, vol. 11, no. 2, pp. 81–88, 2021.

[48] D. Pérez-Jorge, M. del C. Rodríguez-Jiménez, E. Ariño-Mateo, and F. Barragán-Medero, "The effect of covid-19 in university tutoring models,"

Sustainability, vol. 12, no. 20, p. 8631, 2020.

[49] C. Chabbott and M. Sinclair, "SDG 4 and the COVID-19 emergency: Textbooks, tutoring, and teachers," Prospects, vol. 49, no. 1, pp. 51–57, 2020.

[50] K. S. C. Tragodara, "Virtual tutoring from the comprehensive training model to Engineering students during the COVID-19 pandemic," in 2021 IEEE World Conference on Engineering Education (EDUNINE), IEEE, 2021, pp. 1–6.

[51] C. Johns and M. Mills, "Online mathematics tutoring during the COVID-19 pandemic: recommendations for best practices," Primus, vol. 31, no. 1, pp. 99–117, 2021.

[52] T. Button and R. Lissaman, "Using live online tutoring to provide access to higher level Mathematics for pre-university students," in The 10th International Conference on Technology in Mathematics Teaching, 2011, p. 94.

[53] J. Hofmeister, "Evaluation research findings of the pre-university project on transition and student mentoring into University," Mentoring and tutoring by students, pp. 107–117, 1998.

[54] S. E. Volet and P. D. Renshaw, "Cross-cultural differences in university students' goals and perceptions of study settings for achieving their own goals," Higher Education, vol. 30, no. 4, pp. 407–433, 1995.

[55] S. B. Kotsiantis and P. E. Pintelas, "Predicting students marks in hellenic open university," in Fifth IEEE International Conference on Advanced Learning Technologies (ICALT'05), IEEE, 2005, pp. 664–668.

[56] J. L. Arco-Tirado, F. D. Fernández-Martín, and J.-M. Fernández-Balboa, "The impact of a peer-tutoring program on quality standards in higher education," Higher Education, vol. 62, no. 6, pp. 773–788, 2011.

[57] D. J. Goldsmith, D. Nielsen, G. Rezendes, and C. A. Manly, "Basic eSkills–Foundation or Frustration: A Research Study of Entering Community College Students' Computer Competency.," Online Submission, 2006.

[58] S. Rodríguez Espinar and M. Álvarez González, Manual de tutoría universitaria recursos para la acción. Barcelona: Editorial Octaedro : Universitat de Barcelona, Institut de Ciències de l'Educació, 2012.

[59] J. A. Ross, "Teacher Efficacy and the Effects of Coaching on Student Achievement," Canadian Journal of Education / Revue canadienne de l'éducation, vol. 17, no. 1, pp. 51–65, 1992, doi: 10.2307/1495395.

[60] J. D. Baker, S. A. Rieg, and T. Clendaniel, "An Investigation of an after School Math Tutoring Program: University Tutors + Elementary Students = A Successful Partnership," Education, vol. 127, no. 2, pp. 287–293, 2006.

[61] G. Villa Fernández, J. A. Montero Morales, and A. Llauró Moliner, "Educational coaching applied to group tutoring sessions: An experience with first-year engineering students," in Eighth International Conference on Technological Ecosystems for Enhancing Multiculturality, in TEEM'20. New York, NY, USA: Association for Computing Machinery, Oct. 2020, pp. 339–344. doi: 10.1145/3434780.3436588.

[62] S. E. Volet, "Modelling and coaching of relevant metacognitive strategies for enhancing university students' learning," Learning and instruction, vol. 1, no. 4, pp. 319–336, 1991.

[63] D. Fonseca, J. A. Montero, M. Guenaga, and I. Mentxaka, "Data analysis of coaching and advising in undergraduate students. An analytic approach," in International Conference on Learning and Collaboration Technologies, Springer, 2017, pp. 269–280.

[64] A. Llauró, D. Fonseca, E. Villegas, M. Aláez, and S. Romero, "Educational data mining application for improving the academic tutorial sessions, and the reduction of early dropout in undergraduate students," in Ninth International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM'21), 2021, pp. 212–218.

[65] J. Nielsen, "The usability engineering life cycle," Computer, 1992, doi: 10.1109/2.121503.

[66] R. S. Adams and C. J. Atman, "Cognitive processes in iterative design behavior," in FIE'99 Frontiers in Education. 29th Annual Frontiers in Education Conference. Designing the Future of Science and Engineering Education. Conference Proceedings (IEEE Cat. No.99CH37011, Nov. 1999, p. 11A6/13-11A6/18 vol.1. doi: 10.1109/FIE.1999.839114.

[67] B. Gros and E. Durall, "Retos y oportunidades del diseño participativo en tecnología educativa," Edutec. Revista Electrónica de Tecnología Educativa, no. 74, Art. no. 74, Dec. 2020, doi: 10.21556/edutec.2020.74.1761.

[68] British Design Council, "Design Methods for developing services," An introduction to service design and a selection of service design tools, pp. 1–23, 2007.

[69] C. M. Barnum, Ed., "Praise for Usability Testing Essentials," in Usability Testing Essentials, Boston: Morgan Kaufmann, 2011, p. i. doi: 10.1016/B978-0-12-375092-1.00023-4.

[70] M. Esteban, A. Bernardo, and L. Rodríguez-Muñiz, "Persistence in university studies: The importance of a good start," Aula Abierta, vol. 44, p. 1, Dec. 2016, doi: 10.17811/rifie.44.2016.1-6.

[71] F. Araque, C. Roldán, and A. Salguero, "Factors influencing university drop out rates," Computers & Education, vol. 53, no. 3, pp. 563–574, Nov. 2009, doi: 10.1016/j.compedu.2009.03.013.

[72] R. Gilar-Corbi, T. Pozo-Rico, J.-L. Castejón, T. Sánchez, I. Sandoval-Palis, and J. Vidal, "Academic Achievement and Failure in University Studies: Motivational and Emotional Factors," Sustainability, vol. 12, no. 23, Art. no. 23, Jan. 2020, doi: 10.3390/su12239798.

[73] A. Merlino, S. Ayllón, and G. Escanés, "Variables que influyen en la deserción de estudiantes universitarios de primer año. Construcción de índices de riesgo de abandono / Variables that influence first year university students' dropout rates. Construction of dropout risk indexes," Actualidades Investigativas en Educación, vol. 11, no. 2, Art. no. 2, Sep. 2011, doi: 10.15517/aie.v11i2.10189.

[74] J. Escobar-Pérez and Á. Cuervo-Martínez, "Validez de contenido y juicio de expertos: una aproximación a su utilización," Avances en medición, vol. 6, no. 1, pp. 27-36, 2008.

[75] J. L. Rodríguez, El método Delphi: una técnica de previsión para la incertidumbre. Ariel, 1999. Accessed: Mar. 08, 2022. [Online]. Available: https://dialnet.unirioja.es/servlet/libro?codigo=208626

[76] J. C. Almenara and J. B. Osuna, "La utilización del juicio de experto para la evaluación de tic: el coeficiente de competencia experta," Bordón. Revista de Pedagogía, vol. 65, no. 2, Art. no. 2, Jun. 2013.

[77] M. R. Amaya, D. P. da S. S. da Paixão, L. M. M. Sarquis, and E. D. de A. Cruz, "Construção e validação de conteúdo de checklist para a segurança do paciente em emergência," Rev. Gaúcha Enferm., vol. 37, no. spe, 2016, doi: 10.1590/1983-1447.2016.esp.68778.

[78] E. C. Viera, M. T. A. Robles, F. J. G. Fuentes-Guerra, and J. R. Rodríguez, "Diseño de un cuestionario sobre hábitos de actividad física y estilo de vida a partir del método Delphi," E-Balonmano.com: Revista de Ciencias del Deporte, vol. 8, no. 1, Art. no. 1, Feb. 2012.

[79] M. Varela-Ruiz, L. Díaz-Bravo, and R. García-Durán, "Descripción y usos del método Delphi en investigaciones del área de la salud," Investigación en educación médica, vol. 1, no. 2, pp. 90–95, Jun. 2012.

[80] J. C. Almenara and A. I. Moro, "Empleo del método Delphi y su empleo en la investigación en comunicación y educación," Edutec. Revista Electrónica de Tecnología Educativa, no. 48, 2014, doi: 10.21556/edutec.2014.48.187.

[81] M. R. Álvarez and M. Torrado-Fonseca, "El mètode Delphi," REIRE Revista d'Innovació i Recerca en Educació, vol. 9, no. 1, Jan. 2016, doi: 10.1344/reire2016.9.1916.

[82] J. R. G. Fernández, "Análisis del fenómeno del abandono de los estudios en el Curso de Acceso Directo para mayores de 25 años de la UNED," http://purl.org/dc/dcmitype/Text, UNED. Universidad Nacional de Educación a Distancia, 1989. Accessed: Mar. 01, 2022. [Online]. Available: https://dialnet.unirioja.es/servlet/tesis?codigo=42402

[83] A. Constate-Amores, E. Florenciano Martínez, E. Navarro Asencio, and M. Fernández-Mellizo, "Factores asociados al abandono universitario," Educación XX1, vol. 24, no. 1, Nov. 2020, doi: 10.5944/educxx1.26889.

[84] G. de Fanelli and A. M, "Acceso, abandono y graduación en la educación superior argentina," Serie Debate;5,2007, 2007, Accessed: Mar. 01, 2022. [Online]. Available: http://repositorio.cedes.org/handle/123456789/3699

[85] J. La Madriz, "Factors That Promote the Defection of the Virtual Classroom," Orbis, Revista Científica Ciencias Humanas, vol. 12, no. 35, pp. 18–40, 2016.

[86] C. Luck, "Challenges faced by tutors in Higher Education," Psychodynamic Practice, vol. 16, no. 3, pp. 273–287, Aug. 2010, doi: 10.1080/14753634.2010.489386.

[87] A. B. U. Nájera, J. de la Calleja, "Selection of academic tutors in higher education using decision trees," Revista Española de Orientación y Psicopedagogía, vol. 29, no. 1, pp. 108-124, 2018.

[88] A. Casquero-Tomás, J. Sanjuán-Solís, and A. Antúnez-Torres, "School Dropout by Gender in the European Union: Evidence from Spain," Abandono escolar en función del sexo en la Unión Europea: evidencias sobre España, 2012, Accessed: Mar. 01, 2022. [Online]. Available: https://

dadun.unav.edu/handle/10171/27638

[89] OECD, PISA 2018 Results (Volume VI): Are Students Ready to Thrive in an Interconnected World? Paris: Organisation for Economic Co-operation and Development, 2020. Accessed: May 11, 2023. [Online]. Available: https://www.oecd-ilibrary.org/education/pisa-2018-results-volume-vi_d5f68679-en

[90] R. Z. Oré Ortega, "Comprensión lectora, hábitos de estudio y rendimiento académico en estudiantes de primer año de una universidad privada de Lima Metropolitana," Universidad Nacional Mayor de San Marcos, 2012, Accessed: May 11, 2023. [Online]. Available: https://cybertesis.unmsm.edu.pe/handle/20.500.12672/11512

[91] W. Soto and N. Rocha, "Hábitos de estudio: factor crucial para el buen rendimiento académico," Revista Innova Educación, vol. 2, no. 3, 2020, doi: 10.35622/j.rie.2020.03.004.

[92] A. V. Arguilès, "Del abandono de estudios a la reubicación universitaria," Revista de Sociología de la Educación-RASE, vol. 3, no. 2, 2010, doi: 10.7203/RASE.3.2.8705.

[93] M. Esteban, A. Bernardo, E. Tuero, A. Cervero, and J. Casanova, "Variables influyentes en progreso académico y permanencia en la universidad," European Journal of Education and Psychology, vol. 10, no. 2, pp. 75–81, Dec. 2017, doi: 10.1016/j.ejeps.2017.07.003.

[94] M. (Marina) Elias-Andreu, "Los abandonos universitarios: retos ante el espacio europeo de educación superior," 2008, Accessed: Mar. 01, 2022. [Online]. Available: https://dadun.unav.edu/handle/10171/9139

[95] G. B. Alvarado, "Urban mobility and personal safety as factors related to the decision of dropping out from university," Universidad y Sociedad, vol. 13, no. 5, 2021.

[96] J. Hernández Armenteros, J.A. Pérez García, "La Universidad Española en Cifras. 2017-2018," Spain: Conferencia de Rectores de las Universidades Españolas, ISBN: 978-84-09-18182-7, 2020.

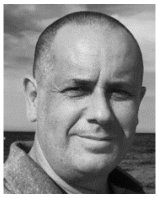
[97] C. Brez, E. M. Hampton, L. Behrendt, L. Brown, and J. Powers, "Failure to Replicate: Testing a Growth Mindset Intervention for College Student Success," Basic and Applied Social Psychology, vol. 42, no. 6, pp. 460–468, Nov. 2020, doi: 10.1080/01973533.2020.1806845.

[98] R. Gilar-Corbi, T. Pozo-Rico, and J. L. Castejón-Costa, "Desarrollando la Inteligencia Emocional en Educación Superior: evaluación de la efectividad de un programa en tres países.," Educación XX1, vol. 22, no. 1, Art. no. 1, 2019, doi: 10.5944/educxx1.19880.

[99] R. Gilar, T. Pozo-Rico, B. Sanchez, and J. L. Castejón, "Promote learning in emotional competence across an e-learning context for higher education," INTED2018 Proceedings, pp. 1374–1380, 2018.

Alba Llauró Moliner

Alba is a teaching assistant in the Engineering department at La Salle Ramon Llull University. She finished her studies in Multimedia Engineering in 2018 and completed her Master in Project Management in 2019, both degrees obtained at La Salle Universidad Ramon Llull, Barcelona, Spain. She is currently a teaching assistant in Design and Usability 1, Multimedia Productions 1 and Multimedia Productions 2. Coordinator of the Medialab lab, motion capture place and TV/Multimedia set. Tutor of first-degree students in the engineering area. Her research activity is developed in the area of teaching innovation, skills development, service-learning and the study of early school dropout, in the research group GRETEL (Group of REsearch on Technology Enhanced Learning). Her research is related to the early dropout of university students in the first year of university studies and how to prevent this possible dropout.



David Fonseca

Full Professor (2017) by La Salle Ramon Llull University, currently he is the coordinator of the Group of Research on Technology Enhanced Learning (GRETEL), a recognized research group of Generalitat de Catalunya (from 2014), and coordinator of the Graphic Representation Area in the Architecture Department of La Salle (where he is a teacher and academic tutor). Technical Engineer in Telecommunications (URL – 1998), Master in GIS (Universitat de Girona, 2003), Audiovisual Communication Degree (UOC, 2006), Master in Advanced Studies (URL-2007), Official Master in Information and Knowledge Society (UOC, 2008), PhD in Multimedia by URL (2011), also, he is Autodesk Approved and Certified Instructor from 1998. With extensive experience in project manager (from 2000 to act, he has coordinated more than 50 local, national, and international projects, both technological transfer and research funded projects): he has directed 7 PhD thesis and more than 10 other final degree and master projects. Currently he is serving as program or scientific committee in more than 15 indexed journals and conferences, as well as organizing workshops, special issues and invited sessions in different scientific forums.

Eva Villegas Portero

Academic coordinator of the Multimedia Engineering Degree, Master in User experience, and professor by La Salle Ramon Llull University. Coordinator of the user experience, accessibility, and gamification research line of the Group of Research on Technology Enhanced Learning (GRETEL) a recognized research group of Generalitat de Catalunya (from 2020). Graduate in Multimedia (UPC-2001), Master in accessible technologies for information society services (UOC-2010), Master in multimedia creation and serious games (URL-2015), PhD from Ramon Llull University (2020). Responsible and consultant for national and international technology transfer projects. Member of scientific committees of national and international journals and conferences.

Marian Aláez Martínez

Marian is an associate professor in the Department of Private Law at the University of Deusto. PhD in Business and Economics, 1998; Master in Advanced Management, 1995 and Specialisation Diploma in University Teaching, 1999. All degrees at the University of Deusto, Spain. She is currently teaching Introduction to Economics and Strategic Management. As a member of the Teaching Innovation Team at the University of Deusto, she has coordinated the accreditation evaluation of teaching staff in both the planning and implementation of teaching, both in undergraduate and master's degree courses. She has trained teaching staff in subject planning, teaching methodologies and the adaptation of soft competences to remote teaching. She has taught and led improvement teams in the framework of new teacher training. She has participated in more than 15 teaching innovation projects, many of which she has led. Her research activity is developed in the area of teaching innovation, competence development, service-learning and the study of early student leaving. Among others, she has participated in international projects such as Tuning India, ALLVET (Russia and Afghanistan) and CALESA (Philippines), co-funded by the Erasmus+ Programme of the European Union and in the institutionalisation project of service-learning UNISERVITATE (Porticus).

Susana Romero Yesa

Susana Romero Yesa is a professor and researcher at the Faculty of Engineering of the University of Deusto since 1997. Degree in Computer Science with specialization in Technical Informatics, 1996; Specialization Diploma in University Teaching, 2003; and PhD in Computer Science, 2015, all degrees from the University of Deusto, Spain. Her teaching activity is focused on Engineering degrees, specifically in subjects related to electronics, as well as in teacher training at the university itself. The scope of his research focuses on engineering education. Thus, his thesis was developed on Learning Analytics in a remote laboratory environment. His first publications address Electronics, Microcontrollers and Microbotics from an educational point of view, and his current research work is mainly focused on educational projects and publications related to the adaptation of degrees to the EHEA and Teaching Innovation, areas in which he has been working at the university since the arrival of the Bologna Process, participating in the creation and dissemination of the UD Pedagogical Framework (2001). Since the academic year 2016-2017, she has been carrying out these activities as a member of the University's Teaching Innovation Unit and Head of Teaching Innovation at the Faculty of Engineering.

A Benchmark for the UEQ+ Framework: Construction of a Simple Tool to Quickly Interpret UEQ+ KPIs

Anna-Lena Meiners[1]*, Martin Schrepp[2], Andreas Hinderks[3], Jörg Thomaschewski[1]

[1] University of Applied Sciences Emden/Leer (Germany)
[2] SAP SE (Germany)
[3] University of Sevilla (Spain)

* Corresponding author: anna-lena.meiners@ux-researchgroup.com

Received 31 May 2022 | Accepted 12 April 2023 | Published 9 May 2023

Abstract

Questionnaires are a highly efficient method to compare the user experience (UX) of different interactive products or versions of a single product. Concretely, they allow us to evaluate the UX easily and to compare different products with a numeric UX score. However, often only one UX score from a single evaluated product is available. Without a comparison to other measurements, it is difficult to interpret an individual score, e.g. to decide whether a product's UX is good enough to compete in the market. Many questionnaires offer benchmarks to support researchers in these cases. A benchmark is the result of a larger set of product evaluations performed with the same questionnaire. The score obtained from a single product evaluation can be compared to the scores from this benchmark data set to quickly interpret the results. In this paper, the first benchmark for the UEQ+ (User Experience Questionnaire +) is presented, which was created using 3.290 UEQ+ responses for 26 successful software products. The UEQ+ is a modular framework that contains a high number of validated user experience scales that can be combined to form a UX questionnaire. Currently, no benchmark is available for this framework, making the benchmark constructed in this paper a valuable interpretation tool for UEQ+ questionnaires.

Keywords

Benchmark, Measurement, UEQ+ Framework, User Experience, UX Questionnaires.

DOI: 10.9781/ijimai.2023.05.003

I. Introduction

User experience (UX) is a key factor for the success of interactive products. It helps to attract new users and fosters loyalty [1]. Loyal customers are less likely to switch to a competitor – or to terminate their contract to switch to a competitor in the case of subscriptions –and more likely to buy a successor product from the same brand or manufacturer. Loyalty takes time to develop, so it is important that a product offers a constantly high level of UX. Thus, it is important to rigorously measure the UX of a product over a longer period of time. UX questionnaires, especially online questionnaires, are a popular measurement method for this purpose [2], since they allow to reach larger groups of users with low effort.

UX covers a large number of different task-related and non-task-related quality aspects concerning the interaction of a user and a product [3]–[5]. For a good UX impression, the product should be easy to learn or even intuitive to use, it should react to the user's commands as she or he expects, have a visually aesthetic user interface, and be fun to use, among other criteria. Therefore, to guarantee a high level of UX, several semantically distinct UX quality aspects must be considered.

How important such distinct UX aspects are for the overall UX impression of a product depends on personal preferences and on demographic attributes of the user – and, even more importantly, on the type of product [6]–[8]. For instance, efficiency is a key UX requirement for a business software that is used frequently in a typical workday. However, an intuitive interaction is not really expected in this case, since complex products typically require some learning phase, a factor which is considered in the design. Things look different for a web service that is used more sporadically, e.g. a tool to book concert tickets online: In this case, efficiency is nice, but it is not a key factor, whereas an intuitive interaction is absolutely required, since users will not accept any learning time for the simple tasks involved. This is a relatively simple example. For a more detailed analysis of the dependency between product type and the importance of UX aspects, see [6]–[8].

Of course, the number of questions that can be used in an online questionnaire is limited. Users cannot be expected to spend much time answering evaluation questions. At the same time, it is critical to measure the most important UX aspects, which will be different for each product. Measuring all the right aspects thus cannot be achieved using standardized UX questionnaires [3], [9], which contain a fixed set of items and scales.

Please cite this article as:
A. L. Meiners, M. Schrepp, A. Hinderks, J. Thomaschewski. A Benchmark for the UEQ+ Framework: Construction of a Simple Tool to Quickly Interpret UEQ+ KPIs, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 1, pp. 104-111, 2024, http://dx.doi.org/10.9781/ijimai.2023.05.003

The UEQ+ is a new modular framework that addresses this issue [9]. It contains a catalog of scales that can be combined to form a concrete UX questionnaire. Thus, a researcher can define the UX aspects that are important for a product and pick the UEQ+ scales that measure those aspects.

Questionnaires built with the UEQ+ framework can already be used to compare different products concerning UX or to measure how the UX of a product develops over time [10]–[11]. However, no benchmark is currently available for the UEQ+. This is problematic since it is difficult to interpret the UX score of a single product without an appropriate comparison.

The goal of this paper is to define a first benchmark for the UEQ+ that will help UX researchers and practitioners to interpret standalone UEQ+ results.

II. The UEQ+ Framework

The UEQ+ [9] is a catalog of UX scales. It extends the UEQ [12]–[13] with additional scales. Each of these scales describes a special semantic aspect of the interaction between a user and a product. As already mentioned in the introduction, the UX aspects measured in a product evaluation depend on the product type and the specific research question. The modular structure of the UEQ+ addresses this requirement. Researchers can pick exactly those scales from the available UEQ+ scales that are most important for their study and can thus set up a questionnaire that optimally fits their research questions.

All UEQ+ scales share the same format, meaning that they can be combined in any order. The following UEQ+ scale shows the standardized format of four items related to a particular UX aspect, perspicuity, followed by a question on the importance of these items for the user:

In my opinion, handling and using the product are:

not understandable	o o o o o o o	understandable
difficult to learn	o o o o o o o	easy to learn
complicated	o o o o o o o	easy
confusing	o o o o o o o	clear

I consider the product property described by these terms as

completely irrelevant	o o o o o o o	very important

As shown above, the items of a UEQ+ scale consist of two terms that represent the opposite ends of a semantic dimension (for example, confusing/clear). Participants can describe their impression on a 7-point answer scale. An introductory sentence is used to set a common context for the four items. Following this, the question at the end is used to calculate an overall UX KPI by weighting the rating of a scale with its importance. The idea behind this KPI is identical to the KPI calculation for the original UEQ, see [14]. Since it is central for this paper, the calculation of the UEQ+ KPI is described in more detail in the following paragraphs.

Assume the scales S_1, …, S_m have been chosen from the available UEQ+ scales. Thus, the final questionnaire contains m scales. Assume further that data has been collected from n participants in the study. Let s_{ij} be the score (the average of the 4 items in the scale) and w_{ij} the importance rating of participant i concerning scale j. Now, firstly, the relative importance r_{ij} of scale j for participant i is calculated by (1).

$$r_{ij} = \frac{w_{ij}}{\sum_{j=1}^{m} w_{ij}} \quad (1)$$

Thus, r_{ij} is the importance rating of scale j by participant i divided by the sum of all importance ratings of participant i. The KPI is then calculated by (2).

$$KPI = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} r_{ij} s_{ij} \quad (2)$$

Thus, the ratings per scale are first weighted with their relative importance per participant, and then the KPI is calculated as the average over the obtained values of all participants. Semantically, the UX KPI represents the overall UX impression that the participants got from the evaluated product. This KPI can easily be calculated in the data analysis Excel tool available on the UEQ+ homepage https://ueqplus.ueq-research.org.

The UEQ+ was developed as an extension of the UEQ and thus contains the 6 scales from the UEQ plus additional scales. The item format is slightly different to allow a free combination of scales. The main difference is that in the original UEQ the introductory sentence is missing, the items are not grouped by scales but appear in random order and that half of the items show the positive term on the left and the other half on the right. The reasons for the changes compared to the original UEQ item format are described in detail in [9].

As of today, the UEQ+ contains 20 UX scales (more scales may be added in the future) that represent different UX aspects of interactive products. Some examples of other scales are:

- *Efficiency:* Can users solve their tasks without unnecessary effort? Does the product react quickly?
- *Visual Aesthetics:* Does the product look beautiful and appealing?
- *Usefulness:* Does using the product bring advantages to the user?
- *Response quality:* Do the responses of a voice assistant satisfy the user's needs of information?
- *Acoustics:* Impact of sounds or operating noise of the product to the user experience.

Some scales are applicable to a large variety of products, for example efficiency, perspicuity, usefulness, or trust. Others make sense only for specific product types, for example acoustics (sound created by operation of a device – which was constructed to evaluate household appliances), aesthetics (only for products with a graphical user interface), or response quality (only for voice assistants).

The complete list of available scales can be found on the UEQ+ homepage https://ueqplus.ueq-research.org. Translations of the scales to more than 30 different languages are also available on this page. The homepage additionally offers supporting material, for example a data analysis Excel tool and a handbook that describes best practices concerning the usage of the UEQ+. The UEQ+ itself and all the provided material on the homepage are free to use.

Creating a concrete UX questionnaire for a product evaluation based on the UEQ+ is simple. The researcher must decide which scales are relevant for the product that should be evaluated. Then these scales are placed in a sequence to form the concrete questionnaire. The recommendation is to use at most 6 scales in a questionnaire constructed from the UEQ+ framework to keep the time required to fill it out within a reasonable range. If more scales are needed to get a clear picture on the UX of a product, it is recommended to split them into two different questionnaires, i.e., to collect data from two samples with a reduced number of scales.

Scales for the UEQ+ framework can be constructed independently and tested for their reliability and validity. This makes it possible to enhance the framework step by step with additional scales [9].

III. Related Research: Different UX Benchmarks

A UX questionnaire allows us to compare products with respect to the scales found in that questionnaire. If product A obtains a significantly higher score in a scale than product B, then A is better than B concerning the UX quality measured by this scale. However, if we have only a single measurement, it is usually difficult to interpret its value directly. For example, is a mean value *Efficiency* = 1,1 on the UEQ+ scale a good or bad result [15]? Does it indicate that the efficiency of the product is sufficiently high? This question can only be answered by comparing the measured score with scores obtained for other products.

This is the idea behind benchmarks. A benchmark is the result of a collection of measurements obtained from different products with a UX questionnaire. Thus, a benchmark allows us to determine how well the evaluated product has performed compared to the products in the benchmark data set. Several standardized questionnaires, for example the User Experience Questionnaire (UEQ) [16], the System Usability Scale (SUS) [17], the Software Usability Measurement Inventory (SUMI) [18], the Usability Metric for User Experience (UMUX) [19], the shorter version UMUX-Lite [20], the Standardized User Experience Percentile Rank Questionnaire (SUPR-Q) [21] or the Visual Aesthetics of Websites Inventory (VisAWI) [22], include such benchmarks.

How exactly a benchmark is defined depends on the questionnaire. This is illustrated by some examples in what follows.

First, we look at the SUS [23] benchmarks. The SUS questionnaire contains 10 items that can be answered on a 5-point Likert scale, with scale values ranging from 0 to 4. The questionnaire does not provide separate scales to measure sub-aspects of the UX. Rather, it offers a single value that represents the overall usability of the product. To calculate this overall value, all 10 item scores of a participant are summed up, resulting in a value between 0 and 40. This value is then multiplied by 2,5, so that the SUS score is stretched to a range from 0 to 100. According to [24], this rescaling was done primarily because a scale of 0 to 100 is easier to communicate to product managers than a scale of 0 to 40. Therefore, the rescaling is done to improve the communication of the results.

Another benchmark is found in [25] and [26], in which a 7-level rating is derived from a very large collection of SUS data. Here, the SUS values per person were compared to an overall assessment of the product. This overall rating was made possible by seven terms, one of which had to be chosen. This allows a relationship between the overall assessment of a person and the SUS score of the person for the evaluated product. The seven terms are listed below. The value behind the term is the average SUS rating associated with the term (average of the SUS ratings from participants that used this term as an overall rating), whereas the numerical value in parentheses is the standard deviation of this rating.

- *Best imaginable*: 90,9 (13,4)
- *Excellent*: 85,5 (10,4)
- *Good*: 71,4 (11,6)
- *OK*: 50,9 (13,8)
- *Poor*: 35,7 (12,6)
- *Awful*: 20,3 (11,3)
- *Worst Imaginable*: 12,5 (13,1)

Let's assume we measure a product with the SUS and get a score of 25. This would correspond to an overall rating between *Awful* and *Poor* (leaning towards *Awful*). If we get an overall score of 87, this corresponds to an overall rating of *Excellent*. Thus, this simple benchmark helps to interpret a single SUS result by relating the SUS score to a statement about overall UX quality.

Another SUS benchmark with more recent data is presented in [23]. This paper provides 11 categories for the results (see Table I) based on a benchmark set of 241 industrial usability studies. The percentile x-y can be interpreted as follows: x percent of the products from the benchmark showed a result lower than your score, 100-y of the products showed a better result.

Thus, if you obtain a score of 25 then you are in category F and your product belongs to the 14% worst products in the benchmark. If you get a score of 85, then your product would be in category A+ and you are amongst the best 4% of products in the benchmark. Again, this benchmark is very helpful to decide if a single measurement obtained for a product indicates a good, average, or bad UX.

TABLE I. SUS Benchmark as Formulated in [23]

Category	Score Interval		Percentile
A+	84,10	100,00	96–100
A	80,80	84,00	90–95
A-	78,90	80,70	85–89
B+	77,20	78,80	80–84
B	74,10	77,10	70–79
B-	72,60	74,00	65–69
C+	71,10	72,50	60–64
C	65,00	71,00	41–59
C-	62,70	64,90	35–40
D	51,70	62,60	15–34
F	0,00	51,60	0–14

These examples show that the goal of a benchmark is to provide some interpretation concerning how good or bad a measured result is overall. How this is formulated in detail is not standardized, but rather decided by the researchers that set up the benchmark.

The benchmark of the UEQ, as described in [16] and [27], works in a similar way to the benchmark described in [23]. It is based on the data of 21.175 participants from 468 different studies, where one study corresponds to one measurement of one product made with the UEQ. For each scale, the measured value is divided into 5 categories:

- ***Excellent:*** The measured scale value is among the top 10% of the best results.
- ***Good:*** The measured scale value is better than 75% of the measured results and worse than the 10% best results.
- ***Above Average:*** The measured scale value is better than 50% of the measured results and worse than the 25% best results.
- ***Below Average:*** The measured scale value is better than 25% of the measured results and worse than the 50% best results.
- ***Bad:*** The measured scale value is among the 25% worst results.

The UEQ benchmark is available in the data analysis Excel tool for the UEQ that can be downloaded from www.ueq-online.org. The graphical representation in Fig. 1 is automatically calculated from the measured scale values of the 6 UEQ scales for a product. For the example product shown in Fig. 1 (black line) we can see that it is highly rated in the pragmatic quality aspects *Perspicuity*, *Efficiency* and *Dependability*, but shows a low rating for aspects that are related to fun of use, i.e., *Stimulation* and *Novelty*. Thus, in this example the result also provides some clear indication on how to improve the product in the future. One could conclude that investments in the usability of the product are not necessarily important, but improvements concerning fun of use will most likely have a big impact on the overall UX.

The VisAWI, as presented in [22] and [28], pursues a somewhat different type of benchmarking. A large amount of existing data from more than 160 different web pages was evaluated. These were additionally divided into categories, for example weblogs and social sharing, e-commerce, and information. For each group, the benchmark shows the mean value and standard deviation for the 4 scales of the VisAWI. A convenient feature of this benchmark is that one can always compare a result with a group of very similar web pages. However, the evaluation is somewhat limited, since one can only conclude whether the VisAWI result for a web page is above or below average.

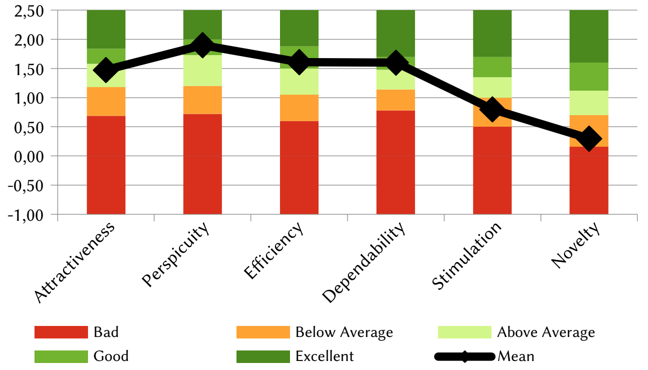


Fig. 1. Comparison of the UEQ measurement of a hypothetical product with the products in the benchmark data set. The UEQ scale ranges from -3 (worst) to 3 (best).

To sum up, there are various ways of creating a benchmark, and existing questionnaires have taken slightly different paths in this respect. However, the aim of a benchmark is always to make the results of a single UX questionnaire easier to interpret, and to do so it is providing a comparison with a large amount of existing data from other product evaluations.

IV. Designing the Study: Preliminary Thoughts on the UEQ+ Benchmark

As was shown in the last section, benchmarks support the interpretation of UX questionnaire results. For the UEQ+, no such benchmark is available so far, and the goal of this paper is to provide at least a first benchmark to fill this gap.

As we have seen, the UEQ+ is a flexible framework that offers a catalog of currently 20 scales to the researcher [9]. The flexibility of such a modular approach has some drawbacks. Providing a benchmark for each scale is extremely difficult since it requires the collection of many product evaluations with the same scale. Since some of the scales are only useful for specific product types, this is very difficult in practice. Thus, providing a benchmark for the single scales, as needed for the UEQ, will require a long time. Since scales will be added from time to time, such a benchmark will also never cover all UEQ+ scales with the same quality.

Therefore, this study follows a different approach: It defines a first simple benchmark based on a limited set of product evaluations, which will provide some quick guidance to UX researchers.

This benchmark is based on the UX KPI, i.e., not on single scales but on the overall user's impression of a product. Of course, this KPI depends on the scales selected for evaluation and this selection varies from product to product. So, it is highly questionable if products from different categories can be reasonably compared by comparing their KPIs. However, from a practical perspective this is not relevant in most cases. Often a comparison to some competitors or products serving the same use cases is sufficient to get a first idea as to whether the own product is "good enough" concerning UX. And since the importance of UX aspects depends on the product type [6]–[8], a similar set of scales will most likely be chosen to measure products of the same type.

Thus, this benchmark was constructed based on one or two representatives for each product type. The KPI of each product was measured via a questionnaire with scales from the UEQ+. For each product, scales were chosen based on how important certain UX aspects are to the users of that product, as reported in previous research [6]–[8].

V. Conducting the Study: Creation of the UEQ+ Benchmark

The data used to create the benchmark was collected in two sets of studies: (1) the data of four studies was taken from our research repository from already existing UEQ+ studies perfectly matching our requirements (inventory data), and (2) the data of 22 additional studies that were collected in two waves, in the fall of 2020 and the fall of 2021 (original data). Details of the data collection are described in the following paragraphs.

Of the inventory data sets, three (regarding Amazon Prime Video, otto.de & zalando.de) were collected as part of validation studies for the UEQ+ [9], and one (regarding Facebook) was part of an interpretation analysis study [29]. The Facebook data was collected via an English social panel. All admissible participants stated that they used Facebook at least once a month.

The questionnaires used for the validation studies were distributed via e-mail and linked on websites. Table II shows participant details for these inventory data sets, specifically their language, mean age and gender attributes.

TABLE II. Participant Details of Inventory Data Sets (Responses, Number of People Stating They Are Male, Female, and Diverse or Not Specified, Mean Age and Language Version of the UEQ+ Used)

UEQ+ Study	N	m	f	d/ not specified	Mean age	Language version
Amazon Prime Video	57	36	21	0	32	German
Facebook	248	112	132	4	30	English
otto.de	42	16	25	1	34	German
zalando.de	46	20	24	2	31	German

Of the original data sets, three were collected via social panels (Alexa, bbc.com & Ebay). The remaining 19 data sets were collected via multi-channel distribution of questionnaires supported by university students at University of Applied Sciences Emden/Leer. Students shared the questionnaires via e-mail, forums, social media, and messengers.

All of these questionnaires were made available in multi-language versions using standardized translations for German and English, so that participants could answer in the language they preferred, and the results could be merged into one database per study. The collected data was then cleaned using the following exclusion criteria:

(1) perfect duplication of entries (oldest entry kept)

(2) time needed to complete questionnaire was below 50 seconds

(3) stated age was over 90 or below 16

(4) less than 80% of UEQ+ items were answered

(5) control questions were answered wrong

Information on the remaining participants in the original studies is found in Table III.

Each questionnaire started with a short introduction followed by demographic questions (gender and age). Then the sequence of UEQ+ items (see Fig. 3 in the Appendix) was displayed. After this block, in most cases, comment fields and questions were added, in order to detect persons that do not answer the questions seriously.

The scales contained in the UEQ+ questionnaires were selected by product. Previous studies [6]–[8] investigated the importance of common UX aspects (corresponding to UEQ+ scales) for typical product types. We used these results for the selection of the scales. Thus, for each investigated product, the corresponding product type was determined and then the most important scales according to the results of these studies were selected. As suggested in the UEQ+

TABLE III. Participant Details of Original Data Sets (Responses, Number of People Stating They Are Male, Female, and Diverse or Not Specified, Mean Age and Language Version of the UEQ+ Used)

UEQ+ Study	N	m	f	d/ not specified	Mean age	Language version
AirBnB	91	39	49	3	30	89 German, 2 English
Alexa	100	55	43	2	27	English
Amazon	208	110	92	6	38	German
bbc.com	98	31	67	0	37	English
Booking.com	49	20	26	3	36	46 German, 3 English
Ebay	100	49	48	3	30	English
Google Maps	111	63	41	7	31	German
Instagram	97	36	56	5	27	German
Moodle	93	39	49	3	31	German
MS Excel	120	53	61	6	34	89 German, 15 English
MS Teams	130	84	30	16	38	German
MS Word	70	42	22	6	38	German
Netflix	46	27	17	2	30	German
Skype	57	26	24	7	37	German
Spotify	245	116	120	9	28	243 German, 2 English
TikTok	51	21	29	1	26	49 German, 2 English
Trello	28	14	12	2	35	27 German, 1 English
Udemy	41	23	17	1	32	40 German, 1 English
WhatsApp	176	72	92	12	32	174 German, 2 English
Wikipedia	444	104	251	89	28	439 German, 5 English
YouTube	517	409	91	17	25	German
Zoom	25	12	11	2	37	German

handbook, a maximum of 6 scales was used in a single study. Table V in the Appendix summarizes which scales were used in the different questionnaires.

VI. Results: The UEQ+ Benchmark

Following exclusion, a total of 3.290 responses to our surveys were used for this study. 1.629 participants identified as male, 1.447 as female, and 214 didn't specify their gender. The vast majority of questionnaires was answered in the German language version (2.700 responses); another 590 questionnaires were answered in the English language version. The mean age of participants was 33 years.

From the collected data a single UX KPI was calculated per product. As explained, these KPIs make up the initial simple benchmark we were striving to introduce in this study.

Table IV shows the measured values for the UX KPI per product. The scale for the UEQ+ KPI ranges from -3 (worst possible) to +3 (best possible). The measured KPIs ranged from +0,39 (for the social network Facebook) to +1,82 (for the navigation software Google Maps).

Fig. 2 displays the confidence intervals of the KPIs for the products in our benchmark data set, providing a quick overview of the results.

VII. Conclusions: Interpreting the UEQ+ Benchmark

The benchmark data (see Table IV and Fig. 2) already allows for a first rough interpretation of UEQ+ KPIs. The products investigated are all well-known, established products, and in many cases, they are the market leaders in their segments. Thus, they have a certain level of UX maturity. Overall, a rating between +1 and +2 for the KPI seems to represent a good level of UX.

TABLE IV: Investigated Product, Mean, Standard Deviation and Confidence Interval of the KPIs and Number of Responses per Study

Product	KPI	Std	95% Conf. Int.		Responses
Google Maps	1,82	0,67	1,70	1,94	111
Wikipedia	1,79	0,63	1,73	1,85	444
zalando.de	1,70	0,69	1,50	1,90	46
Spotify	1,66	0,76	1,56	1,76	245
Udemy	1,66	0,71	1,44	1,88	41
YouTube	1,60	0,67	1,54	1,66	517
BBC.com	1,54	0,89	1,36	1,72	98
Zoom	1,47	0,85	1,14	1,80	25
MS Excel	1,46	0,89	1,30	1,62	120
Alexa	1,46	0,74	1,31	1,61	100
Netflix	1,43	0,86	1,18	1,68	46
Booking.com	1,41	0,83	1,18	1,64	49
WhatsApp	1,39	0,87	1,26	1,52	176
Amazon Prime Video	1,35	0,87	1,12	1,58	57
Trello	1,29	0,67	1,04	1,54	28
otto.de	1,27	0,90	1,00	1,54	42
Ebay	1,25	0,95	1,06	1,44	100
Amazon	1,25	0,82	1,14	1,36	208
MS Teams	1,18	0,92	1,02	1,34	130
MS Word	1,03	0,96	0,81	1,25	70
Instagram	0,97	0,84	0,80	1,14	97
AirBnB	0,96	0,99	0,76	1,16	91
Moodle	0,71	0,88	0,53	0,89	93
Skype	0,60	1,05	0,33	0,87	57
TikTok	0,54	0,95	0,28	0,80	51
Facebook	0,39	1,06	0,26	0,52	248

Some examples can give an idea of how this benchmark can be used and interpreted more in depth. Assume you are a UX researcher and in your recent project you evaluate a web shop. The measured KPI is around 1,5. Is this an indicator of a good or of a poor UX impression? If you look into the list of evaluated products, you find two popular German web shops (otto.de and zalando.de). Their 95% confidence intervals are 1,00 to 1,54 for otto.de and 1,50 to 1,90 for zalando.de. Thus, the UX of your shop is between these two established shops, which clearly indicates that your UX is most likely in an acceptable range.

Assume now that you evaluate a new messenger. The product is new, offers some special services, but must still compete against established products to gain market share. Thus, it makes sense to check if the UX impression of the new messenger is at least close to the UX impression of *WhatsApp* (whose KPI has a confidence interval of 1,26–1,52. Thus, if you obtain a 0,5 score you immediately see that you cannot compete in terms of UX. On the other hand, if you scored 1,4, you would be in a comparable range to the UX of *WhatsApp*.

This benchmark is helpful even in cases where there is no direct match between the products in this benchmark and the evaluated product. Since these are all common products, it is easy to get a personal impression of the UX of these products. So, even a simple first-glance-statement such as "our product generates a UX impression similar to Ebay" can help interpret the results semantically.

In conclusion, with the help of this simple benchmark, UEQ+ results can be compared quickly and easily, which aids practitioners and UX researchers alike by giving an orientation on how to interpret the results of the UEQ+.

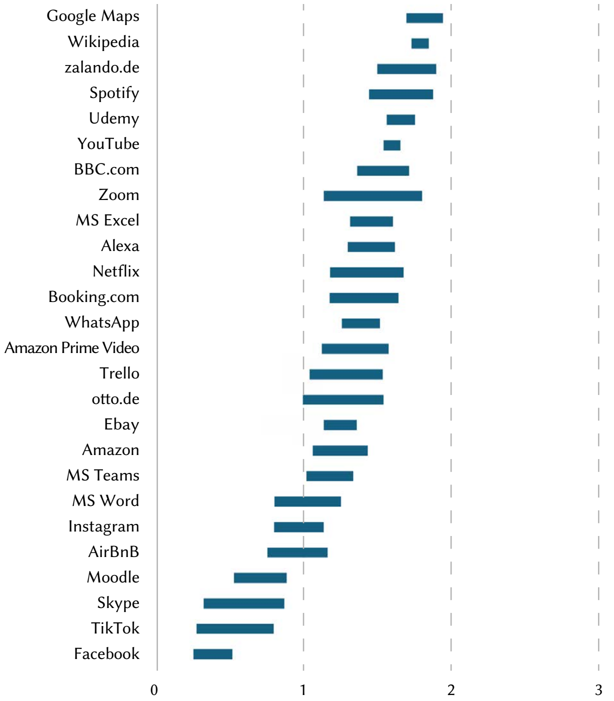


Fig. 2. Confidence intervals for the UEQ+ KPIs of the evaluated products, sorted by KPI.

VIII. Summary and Outlook

A first benchmark for the UEQ+ framework was created. Due to the modular structure of the UEQ+, it is difficult to provide a classical benchmark on the level of single scales. Indeed, since some of the scales are only useful for specific products, it would require a long time to collect enough data for such a benchmark, as with the benchmark of the original UEQ.

Therefore, this first benchmark is based on the UEQ+ KPI and on the evaluation of several well-known products by over 3.200 study participations. A comparison to the KPI values of these products enables UX researchers to develop a first, quick understanding of how good the UX of the evaluated product is compared to user expectations.

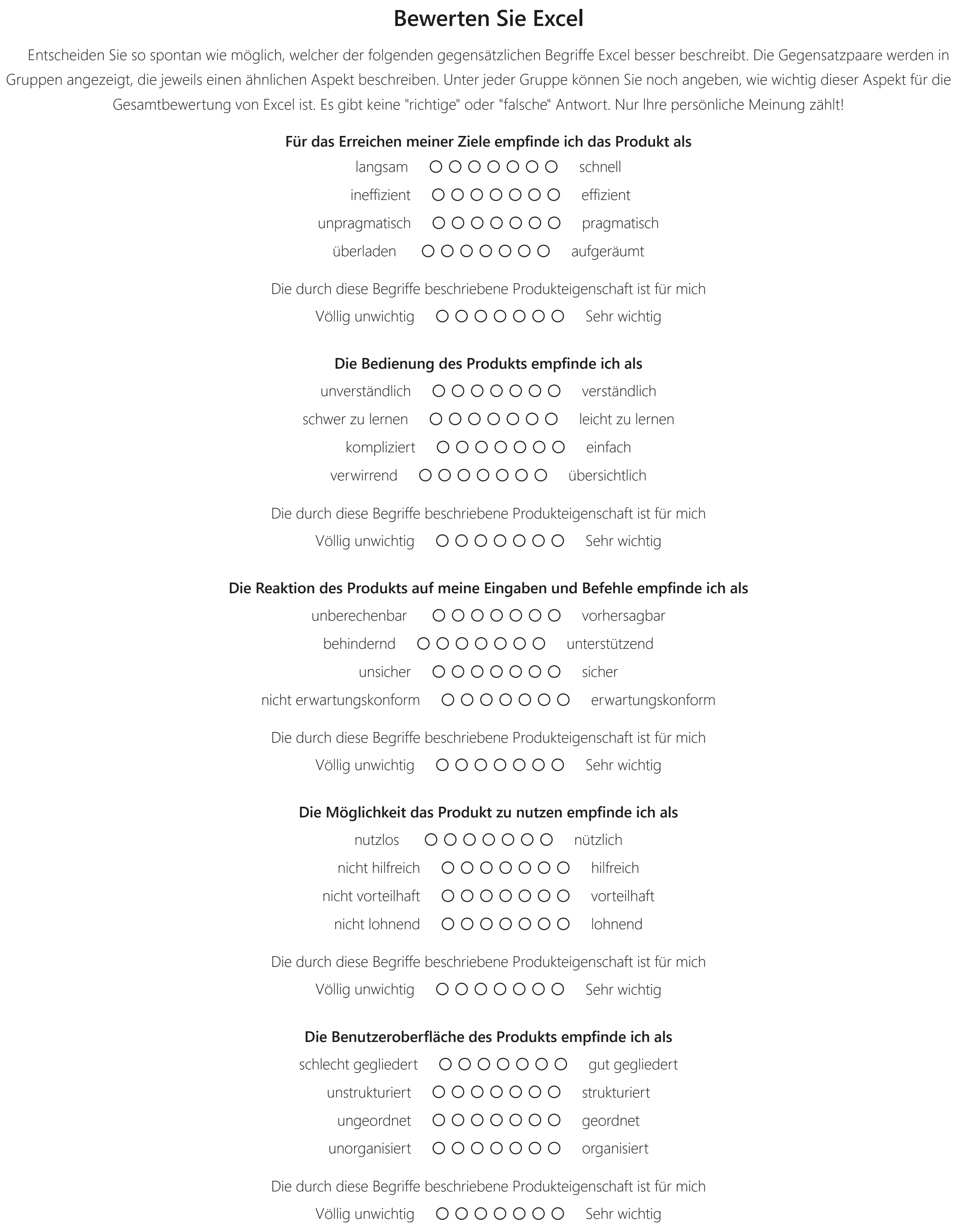
Future lines of work should include studies that give insight into the comparability of products through their UEQ+ KPIs. As mentioned before, it is questionable whether products from different categories are comparable through their UEQ+ KPI. This should be investigated further as it would result in knowledge useful for an easier interpretation of UX data. Further studies should also include more transparent and/or standardized ways of choosing UEQ+ scales when creating the questionnaires as this seems to have happened almost arbitrarily at some points in the past. This would benefit the replicability of the studies and enable expedient comparability of results, which in turn allows for more meaningful and nuanced interpretation of a KPI benchmark. Some research supporting UEQ+ users in this regard has already been conducted [6]–[8], but should be expanded on. It is also possible that product categories such as "banking software" and "shops" don't partition the available data in the most insightful way regarding the comparability of the KPIs, and that instead more abstract categories such as "software for work" or "tools people use rarely" would generate more useful insights. Studies in this regard could begin to create a model of product categories that categorize products from a UX perspective.

Nevertheless, this first benchmark provides a valuable insight for UX practitioners to judge the UX quality of the products they design and evaluate.

Appendix

TABLE V. Investigated Products and Scales Used in the Study

Product	UEQ+ Scales
Facebook	Quality of Content, Trustworthiness of Content, Intuitive Use, Trust, Stimulation
Alexa	Response behavior, Response quality, Comprehensibility, Trust, Usefulness
BBC.com	Perspicuity, Value, Intuitive Use, Quality of Content, Clarity
zalando.de	Attractiveness, Dependability, Intuitive Use, Visual Aesthetic, Quality of Content, Trustworthiness of Content, Trust, Value
Ebay	Trust, Quality of Content, Dependability, Clarity, Intuitive Use
Google Maps	Efficiency, Perspicuity, Usefulness, Intuitive Use, Trustworthiness of Content, Quality of Content
Instagram	Attractiveness, Stimulation, Novelty, Trust, Visual Aesthetic, Intuitive Use, Clarity
Netflix	Attractiveness, Perspicuity, Stimulation, Visual Aesthetics, Intuitive Use, Quality of Content
Teams	Efficiency, Perspicuity, Dependability, Trust, Usefulness, Clarity
Trello	Efficiency, Perspicuity, Trust, Adaptability, Usefulness, Clarity
WhatsApp	Efficiency, Perspicuity, Dependability, Trust, Intuitive Use, Clarity
Word	Efficiency, Perspicuity, Dependability, Usefulness, Intuitive Use, Clarity
YouTube	Attractiveness, Perspicuity, Dependability, Stimulation, Intuitive Use, Clarity
Amazon Prime Video	Attractiveness, Perspicuity, Intuitive Use, Visual Aesthetic, Quality of Content, Trustworthiness of Content, Trust
Moodle	Attractiveness, Perspicuity, Dependability, Adaptability, Usefulness, Clarity
otto.de	Attractiveness, Dependability, Intuitive Use, Visual Aesthetic, Quality of Content, Trustworthiness of Content, Trust, Value
AirBnB	Trust, Quality of Content, Dependability, Efficiency, Clarity
Amazon	Trust, Quality of Content, Dependability, Clarity, Intuitive Use
Booking.com	Trust, Quality of Content, Dependability, Efficiency, Clarity
Excel	Usefulness, Dependability, Efficiency, Perspicuity, Clarity
Skype	Trust, Dependability, Efficiency, Usefulness, Intuitive Use
Spotify	Perspicuity, Dependability, Stimulation, Adaptability, Intuitive Use
Tiktok	Trust, Dependability, Intuitive Use, Quality of Content, Stimulation
Udemy	Quality of Content, Usefulness, Clarity, Perspicuity, Efficiency
Wikipedia	Quality of Content, Clarity, Perspicuity, Visual Aesthetic, Intuitive Use
Zoom	Trust, Dependability, Efficiency, Usefulness, Intuitive Use



Bewerten Sie Excel

Entscheiden Sie so spontan wie möglich, welcher der folgenden gegensätzlichen Begriffe Excel besser beschreibt. Die Gegensatzpaare werden in Gruppen angezeigt, die jeweils einen ähnlichen Aspekt beschreiben. Unter jeder Gruppe können Sie noch angeben, wie wichtig dieser Aspekt für die Gesamtbewertung von Excel ist. Es gibt keine "richtige" oder "falsche" Antwort. Nur Ihre persönliche Meinung zählt!

Für das Erreichen meiner Ziele empfinde ich das Produkt als

langsam ○ ○ ○ ○ ○ ○ ○ schnell
ineffizient ○ ○ ○ ○ ○ ○ ○ effizient
unpragmatisch ○ ○ ○ ○ ○ ○ ○ pragmatisch
überladen ○ ○ ○ ○ ○ ○ ○ aufgeräumt

Die durch diese Begriffe beschriebene Produkteigenschaft ist für mich
Völlig unwichtig ○ ○ ○ ○ ○ ○ ○ Sehr wichtig

Die Bedienung des Produkts empfinde ich als

unverständlich ○ ○ ○ ○ ○ ○ ○ verständlich
schwer zu lernen ○ ○ ○ ○ ○ ○ ○ leicht zu lernen
kompliziert ○ ○ ○ ○ ○ ○ ○ einfach
verwirrend ○ ○ ○ ○ ○ ○ ○ übersichtlich

Die durch diese Begriffe beschriebene Produkteigenschaft ist für mich
Völlig unwichtig ○ ○ ○ ○ ○ ○ ○ Sehr wichtig

Die Reaktion des Produkts auf meine Eingaben und Befehle empfinde ich als

unberechenbar ○ ○ ○ ○ ○ ○ ○ vorhersagbar
behindernd ○ ○ ○ ○ ○ ○ ○ unterstützend
unsicher ○ ○ ○ ○ ○ ○ ○ sicher
nicht erwartungskonform ○ ○ ○ ○ ○ ○ ○ erwartungskonform

Die durch diese Begriffe beschriebene Produkteigenschaft ist für mich
Völlig unwichtig ○ ○ ○ ○ ○ ○ ○ Sehr wichtig

Die Möglichkeit das Produkt zu nutzen empfinde ich als

nutzlos ○ ○ ○ ○ ○ ○ ○ nützlich
nicht hilfreich ○ ○ ○ ○ ○ ○ ○ hilfreich
nicht vorteilhaft ○ ○ ○ ○ ○ ○ ○ vorteilhaft
nicht lohnend ○ ○ ○ ○ ○ ○ ○ lohnend

Die durch diese Begriffe beschriebene Produkteigenschaft ist für mich
Völlig unwichtig ○ ○ ○ ○ ○ ○ ○ Sehr wichtig

Die Benutzeroberfläche des Produkts empfinde ich als

schlecht gegliedert ○ ○ ○ ○ ○ ○ ○ gut gegliedert
unstrukturiert ○ ○ ○ ○ ○ ○ ○ strukturiert
ungeordnet ○ ○ ○ ○ ○ ○ ○ geordnet
unorganisiert ○ ○ ○ ○ ○ ○ ○ organisiert

Die durch diese Begriffe beschriebene Produkteigenschaft ist für mich
Völlig unwichtig ○ ○ ○ ○ ○ ○ ○ Sehr wichtig

Fig. 3: Example of the UEQ+ part of the survey (German).

References

[1] M. Thüring, and S. Mahlke, ”Usability, aesthetics and emotions in human–technology interaction,” *International journal of psychology*, vol. 42, no. 4, 2007, pp. 253-264.

[2] J. R. Lewis, and J. Sauro, “Usability and user experience: Design and evaluation,” *Handbook of Human Factors and Ergonomics*, 2021, pp. 972-1015.

[3] M. Schrepp, *User Experience Questionnaires: How to use questionnaires to measure the user experience of your products?*, KDP, 2021, ISBN-13: 979-8736459766.

[4] J. Preece, Y. Rogers, and H. Sharpe, “Interaction design: Beyond human-computer interaction,” Wiley, New York, 2002.

[5] M. Hassenzahl, “The effect of perceived hedonic quality on product appealingness,” *International Journal of Human-Computer Interaction*, vol. 13, no. 4, 2001, pp. 481-499.

[6] A.-L. Meiners, J. Kollmorgen, M. Schrepp, and J. Thomaschewski, “Which UX aspects are important for a software product? Importance ratings of UX aspects for software products for measurement with the UEQ+,” in *Proceedings of Mensch und Computer 2021 (MuC ‘21)*, Association for Computing Machinery, New York, NY, USA, 2021, pp. 136−139.

[7] H. B. Santoso and M. Schrepp, “The impact of culture and product on the subjective importance of user experience aspects,” *Heliyon*, vol. 5, no. 9, 2019, doi: 10.1016/j.heliyon.2019.e02434.

[8] M. Schrepp, J. Kollmorgen, A.-L. Meiners, A. Hinderks, D. Winter, H. B. Santoso, J. Thomaschewski. On the Importance of UX Quality Aspects for Different Product Categories, International Journal of Interactive Multimedia and Artificial Intelligence, (2023), http://dx.doi.org/10.9781/ijimai.2023.03.001.

[9] M. Schrepp, and J. Thomaschewski, “Design and Validation of a Framework for the Creation of User Experience Questionnaires,” *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 7, 2019, doi: 10.9781/ijimai.2019.06.006.

[10] H. Sandkühler, M. Schrepp, and J. Thomaschewski, “UX Messung mithilfe des UEQ+ Frameworks (Measuring UX with the UEQ+ framework)”, in *Mensch und Computer 2020 - Workshopband*, Gesellschaft für Informatik, Bonn, 2020.

[11] M. Schrepp, H. Sandkühler, and J. Thomaschewski, “How to create short forms of UEQ+ based questionnaires?,” in *Mensch und Computer 2021-Workshopband*, 2021.

[12] B. Laugwitz, T. Held, and M. Schrepp, “Construction and evaluation of a user experience questionnaire,” in *Symposium of the Austrian HCI and usability engineering group*, Springer, Berlin, Heidelberg, 2008, pp. 63-76.

[13] B. Laugwitz, M. Schrepp, and T. Held, “Konstruktion eines Fragebogens zur Messung der User Experience von Softwareprodukten (Construction of a questionnaire to measure UX of software products),” in *Mensch und Computer 2006,* Oldenbourg Wissenschaftsverlag, 2006, pp. 125-134.

[14] A. Hinderks, M. Schrepp, F. J. D. Mayo, M. J. Escalona, and J. Thomaschewski, “Developing a UX KPI based on the user experience questionnaire,“ *Computer Standards & Interfaces*, vol. 65, 2019 pp. 38-44.

[15] M. Schrepp, A. Hinderks, and J. Thomaschewski, “Applying the user experience questionnaire (UEQ) in different evaluation scenarios,” in *International Conference of Design, User Experience, and Usability*, Springer, Cham, 2014, pp. 383-392.

[16] M. Schrepp, A. Hinderks, and J. Thomaschewski, ”Construction of a benchmark for the User Experience Questionnaire (UEQ*),*” *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 4, no. 4, 2017, pp. 40-44.

[17] J. Brooke, “SUS-A quick and dirty usability scale,” *Usability Evaluation in Industry*, vol. 189, no. 194, 1996, pp. 4-7.

[18] J. Kirakowski, and M. Corbett, “SUMI: The software usability measurement inventory,” *British Journal of Educational Technology*, vol. 24, no. 3, 1993, pp.210-212.

[19] K. Finstad, “The usability metric for user experience,” *Interacting with Computers*, vol. 22, no. 5, 2010, pp. 323-327.

[20] J. R. Lewis, B. Utesch, and D. E. Maher, D. E., “UMUX-LITE: when there’s no time for the SUS,” *in Proceedings of the SIGCHI conference on human factors in computing systems*, 2013, pp. 2099-2102.

[21] J. Sauro, “SUPR-Q: A comprehensive measure of the quality of the website user experience,” *Journal of usability studies*, vol. 10, no. 2, 2015.

[22] M. Moshagen, and M. Thielsch, “Facets of visual aesthetics,” *International Journal of Human-Computer Studies*, vol. 68, no. 10, 2010, pp. 689-709.

[23] J. Sauro, and J. R. Lewis, *Quantifying the user experience: Practical statistics for user research*, 2nd ed., Cambridge, MA: Morgan-Kaufmann, 2016.

[24] J. Brooke, “SUS A Retrospective,” *Journal of Usability Studies*, vol. 8, no. 2, 2013, pp. 29-40.

[25] A. Bangor, P. T. Kortum, and J. T. Miller, “An empirical evaluation of the System Usability Scale,” *International Journal of Human-Computer Interaction*, vol. 24, no. 6, 2008, pp. 574-594.

[26] A. Bangor, P. T. Kortum, and J. T. Miller, J. T., “Determining what individual SUS scores mean: Adding an adjective rating scale,” *Journal of Usability Studies,* vol. 4, no. 3, 2009 pp. 114-123.

[27] M. Schrepp, S. Olschner, and U. Schubert, „User Experience Questionnaire (UEQ) Benchmark,” in *Tagungsband UP13*, 2013.

[28] M. Moshagen and M. Thielsch, “A short version of the visual aesthetics of websites inventory,” *Behaviour & Information Technology*, vol. 32, no. 12, 2013, pp. 1305-1311.

[29] A. Hinderks, F. J. Domínguez-Mayo, A.-L. Meiners, and J. Thomaschewski, “Applying Importance-Performance Analysis (IPA) to interpret the results of the User Experience Questionnaire (UEQ),” *Journal of Web Engineering*, vol. 19, no. 2, 2020, pp. 243-266.

Anna-Lena Meiners

Anna-Lena Meiners received a Bachelor’s degree in Theatre Studies, Philosophy and Dutch Language and Literature from Freie Universität Berlin and a Bachelor’s as well as a Master’s degree in Computer Science and Digital Media from University of Applied Sciences Emden/Leer with a focus on Human-Computer Interaction. Currently, she is a PhD researcher at Karlsruhe Institute of Technology (KIT). Her research focusses on technology design for positive and enriching user experiences.

Martin Schrepp

Martin Schrepp has been working as a user interface designer and researcher for SAP SE since 1994. He finished his diploma in mathematics in 1990 at the University of Heidelberg (Germany). In 1993 he received a PhD in Psychology (also from the University of Heidelberg). His research interests are the application of psychological theories to improve the design of software interfaces, the application of *Design for All* principles to increase accessibility of business software, measurement of usability and user experience, and the development of general data analysis methods. He has published several papers concerning these research fields.

Andreas Hinderks

Andreas Hinderks holds a PhD in Computer Science by University of Sevilla. He has worked in various management roles as a Business Analyst and a programmer from 2001 to 2016. His focus lay on developing user-friendly business software. Currently, he is a freelancing Product Owner, Business Analyst and Senior UX Architect. He is involved in research activities dealing with UX questionnaires, measuring user experience and User Experience Management since 2011.

Jörg Thomaschewski

Jörg Thomaschewski received a PhD in physics from the University of Bremen (Germany) in 1996. He became Full Professor at the University of Applied Sciences Emden/Leer (Germany) in September 2000. His research interests are human-computer interaction, agile software development, and e-learning. Dr. Thomaschewski is the founder of the research group “Agile Software Development and User Experience”.

How Does the Visualization Technique Affect the Design Process? Using Sketches, Real Products and Virtual Models to Test the User's Emotional Response

María Alonso-García[1]*, Almudena Palacios-Ibáñez[2], Óscar D. de-Cózar-Macías[3], Manuel D. Marín-Granados[3]

[1] Department of Mechanical Engineering and Industrial Design, University of Cádiz, Avenida de la Universidad 10, 11519, Puerto Real, Cádiz (Spain)
[2] IUI en Tecnología Centrada en el Ser Humano, Universitat Politècnica de València, Camino de Vera, s/n, 46022, Valencia (Spain)
[3] Department of Graphic Expression, Design and Projects, University of Málaga, Arquitecto Francisco Peñalosa, 6, 29071, Málaga (Spain)

* Corresponding author: maria.alonso@uma.es

Received 16 September 2022 | Accepted 10 October 2023 | Published 12 January 2024

unir
LA UNIVERSIDAD
EN INTERNET

Abstract

Testing products during the design process can help design teams anticipate user needs and predict a positive emotional response. Emerging technologies, e.g., Virtual Reality (VR), allow designers to test products in a more sophisticated manner alongside traditional approaches like sketches, photographs or physical prototypes. In this paper, we present the results of a study conducted to evaluate the feasibility of seven visualization techniques for product assessment within the framework of emotional design, suggesting that the user's perception depends on the visualization technique used to present the product. This research provides recommendations for product evaluation using physical, virtual, or conceptual prototypes to analyze the user's emotional response throughout 19 parameters. Our results indicate that the use of virtual environments, including VR and VR with Passive Haptics (VRPH), can facilitate user participation in the design process, although these visualization techniques may also exaggerate the emotions perceived by users. In this context, VRPH tends to overstate the tactile perception of the product. Additionally, our results reveal that both virtual and conceptual environments can amplify a user's likelihood to purchase a product. However, the latter setting could also potentially lead to confusion among users in regards to their perception of the product's weight, dimensions, and cost. Based on these findings, the authors encourage industrial designers to develop new methodologies to optimize design process and minimize costs.

Keywords

Design Process, Emotional Design, Product Evaluation, Virtual Reality.

DOI: 10.9781/ijimai.2024.01.001

I. Introduction

The ongoing evolution of technology and society has led to a greater diversity of products on the market [1]. This has resulted in a highly competitive market with numerous products capable of satisfying the user's basic needs [2], [3]. Therefore, different brands are seeking to bring a differentiating value to their products through innovation to make their products preferred by the user over those of competing brands [4]. However, Marquis and Deeb [5] have shown that this is not always enough, as many innovative products fail when they reach the market despite appearing potentially successful. This stems from the fact that it is difficult for users to analyze a product features due to the large number of alternatives and limited time available [6]. Consequently, the user often base their decision on their instantaneous perception of the product. For this reason, the immediate emotional value that the user attaches to a product is being recognized as an effective differentiation tool [7] and some authors even suggest that these emotions need to be positive for the product to be desired [8]. The creation of a "desirable" and "successful" product is a complex task that relies, among other factors, on testing and user involvement in the design process, especially in the early stages of this process [9]. It is evident that industrial designers must guide users towards a positive emotional response. This perception, as Norman [10] suggests, should be assessed through emotional design before, during, and after product use, at three levels: visceral (VL), behavioral (BL), and reflective (RL). This approach continues to be employed in the design process based on the evaluation of users' perceptions of product features, and it is supported by current studies, specifically Zhe [11] and Aftab et al. [12], who were focused on the analysis of only one or two levels, as well as Göremann and Spiekermann [13], Amirkhizi et al. [14], Yoon et al. [15], and Aftab and Rusli [16].

Please cite this article as: M. Alonso-García, A. Palacios-Ibáñez, Ó. D. de-Cózar-Macías, M. D. Marín-Granados. How Does the Visualization Technique Affect the Design Process? Using Sketches, Real Products and Virtual Models to Test the User's Emotional Response, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 1, pp. 112-126, 2024, http://dx.doi.org/10.9781/ijimai.2024.01.001

Others, as Yu Zhe [11] examined the perception that users had of different ceramic pieces through all these levels. Aftab et al. [12], for example, used these three levels to analyze several disposable products and proposed a design methodology that promoted a long-lasting user-product relationship. On the other hand, Görnemann [13] studied user-product conversations and the resulting person-object relationships, while Jourabchi et al. [14] analyzed the emotional response of users from different cultures to objects with different basic geometric shapes at the VL and identified differences between individualistic and collective societies. Additionally, Yoon et al. [15] focused on the same level and proposed new methodologies for designers to intentionally facilitate positive emotional responses in the user. Bustamante et al. [16] evaluated users' perception of wireless headphones during and after use, focusing on the visceral and behavioral levels.

From the evaluation of products qualities, industrial designers can predict how users will respond once they purchase the product, and the level of success or error achieved during the design process [17]. However, these evaluations are complex due to the number of variables involved in the user's perception, which in many cases depend on psychological factors unknown to the consumers themselves [11].

In this regard, it is important to highlight the difference between the VL and RL in comparison with the BL, even though all of them are related to human, cultural, and psychological factors. According to Chapman [18], while the first two evaluate the user's perception of the product, the BL validates the functional aspects . This author states that the VL and RL are affected by an "emotional" perception of the product, related respectively to the user's immediate exposure to the product and to the reflection and memories perceived afterward. However, the BL aims to study the direct user-object interaction objectively and without depending on the feelings that the product evokes in the user.

Despite these differences, BL is also linked to the user's previous experience with other products and cultural factors. In this context, the suitability of a new product's functionality for the user will depend on their previous experience and how its functionality compares to products they have used before [19]. Several authors as Liberman and Bitan [4], Alonso et al. [17], Liu et al. [20] and Boru and Erin [21] have analyzed this relationship using different parameters .

Currently, the most used are ease of learning, effectiveness, efficiency, memorization, and satisfaction. These parameters were established by Nielsen [22] in 1993 and proposed again by Min and Jeong in 2016 [23], who also proposed the same parameters to evaluate the usability of products. In this regards, recent studies have concluded that the results of any evaluation depend on the user's cultural factors and the formal and technological aspects of the product. These challenges (coupled with the need for multiple physical prototypes during the design process), which can be time-consuming and costly, mean that product evaluation is not used by all design teams [24].

To save costs and shorten the design process in product evaluation from a real environment (RE), some design teams use a conceptual environment (CE). This is accomplished using photorealistic images, renders, virtual models, etc. In this context, Ribelles et al. [25] analyzed the playability of a cubic puzzle game by evaluating parameters such as effectiveness and ease of use through simple visual modifications. Furthermore, the emergence of new visualization methods e.g., Virtual Reality (VR) and Augmented Reality (AR), offers the possibility of presenting products in a virtual environment (VE). These visualization techniques have proven to make the design process more economically efficient, but it is important to consider their limitations.

Hannah et al. [26] have shown that the representation method affects user's perception of the product and can result in perceptual differences from the real product . It is generally assumed that our perceptual and emotional response to a product perceived using different visual media is comparable to that of the physical product, but it is not always the case [27]. Specifically, in product design processes, most studies emphasize media that promote the sense of touch, as it offers a direct way to interact with an object and can minimize these differences [28]. In this context, it is important to determine which visualization technique is the most effective for analyzing different product parameters to obtain accurate user evaluations (compared to the real product) that will not lead to errors during the design process.

This paper explores, under the paradigm of emotional design, different product representation methods as well as the differences in the user's perception that each one presents. This research aims to establish which visualization technique (whether developed in a RE, CE or VE) is best suited for product evaluation during the design process, as this information can prevent future retooling and engineering changes, reduce costs, and ensure the process and everyday product quality. This paper presents (1) an introduction section with the goals of the research, (2) a theoretical background that explores the different media used for product evaluation and their possibilities and limitations, (3) the methodology used for the analysis of the selected visualization techniques, (4) the results and discussion obtained from the data analysis, and (5) the conclusions of the contribution, including advice on the use of different mediums for the evaluation of different parameters during the design process.

II. Theoretical Background

Design professionals often require the construction of physical prototypes [19] to test their designs in a RE. These prototypes are used for product validation [29], the creation of evaluation methodology [30], or product optimization [17].

The CEs seek to represent the product in 2D using photographs, drawings, or photorealistic images. An example of this is the FULE methodology [4], which evaluates products based on their photographic image using the criteria of functionality, usability, and look and feel. In a VE, a 3D representation of the model is used, which can be entirely virtual or a combination of virtual and real elements.

In this context, some researchers have studied the application of VR in engineering and product design [31]. This study examines the feasibility of using VR in different phases of the design process with new visualization technologies as a tool for participatory design. Katicic et al. [32] have developed a specific methodology to evaluate the emotional response of potential customers or consumers to future products during early conceptual design phases. To do this, they created a 6-phase methodology that integrates VR with emotion recognition technologies, allowing users to receive reliable emotional feedback on virtual products in the early stages of product development. There are also studies that investigate different tools for analyzing user-product interaction through VR, but they typically focus on interfaces or work situations, as demonstrated by Gorski et al. [33]. These researchers have implemented a digital tool based on VR to aid the decision-making process in configuring the driver's workstation in urban buses by studying human-machine interaction.

It is important to note that there are discrepancies in the use of the VE for product evaluation. While Liu et al. [20] consider virtual media to be a reliable and cost-effective alternative to physical prototypes, Gorski et al. [33] prefer the use of the RE or a CE. Specifically, Laing and Aperly [34] conducted a study on the opinions of industrial designers and concluded that professionals do not consider virtual media to be an efficient tool. This may be due to the limitations of conceptual and virtual environments, which would require validation of the method used or the characteristics to be measured to determine whether one medium is more appropriate than another [35].

Traditional visualization techniques used in CE allow the user to visually appreciate a product but not to touch it. Similarly, this happens in VR. To address this, new techniques have emerged to provide tactile feedback in VEs. VR with Passive Haptics (VRPH) provides a tactile experience in the VE by superimposing a virtual model on a lower quality physical prototype. In this regard, the differences may be minimized, but the perception of the product may still be affected [36].

Higuera et al. [37] have analyzed the differences between AR and other conceptual and virtual media that prevent touch perception, including photographs, 360° panoramas, and VR. These studies highlight the differences between VEs and CEs. For example, 360° panoramas provide results that are closer to what is perceived in AR according to the psychological responses of participants. VR, on the other hand, obtains higher matches according to physiological responses, which may indicate similarities between virtual and physical interaction.

In a study conducted by Palacios-Ibáñez et al. [38], different users evaluated three coffee makers seen through real photographs and virtual media. The study concluded that the use of immersive media favors the purchase decision and provides greater certainty in the user's response. Furthermore, the results showed that Jordan's sociological pleasure category is more susceptible to media switching in aesthetically rich products. This suggests that users may be more interested in a product in a VE than, which can be a disadvantage when evaluating products as some characteristics may be exaggerated. Therefore, it is necessary to establish measurement systems that take this factor into account. Kent et al. [39] also support the differences between virtual and physical prototypes, validating the use of each in five different dimensions.

According to the conclusions of the refd authors, the choice of the visualization technique is crucial not only for validating the product but also for favoring the purchase decision and ensuring a correct and continuous evaluation of the product during the design process. Additionally, when selecting a method, it is important to consider the characteristics of the product to be evaluated. Not all representation media are valid for all measurements, but all may be useful for measuring specific parameters involved in the design process.

In this context, the emotional value or perception of a product is influenced by both objective and subjective factors (regardless of the visual media), and the latter can also influence the former [40]. Schrepp et al. [41] have pointed out that the aesthetics and usability of products are influential factors in their evaluation. This study suggests that aesthetics can influence the usability of products, or at least the user's perception of the product as Wiedmann et al. [42] done. Some authors even claim that the visual appeal and aesthetics of a product are more important than its functionality and usability [4]. Although it is not yet clear which psychological mechanism is responsible for this relationship [43], using non-functional mock-ups for everyday products makes it easier for users to identify problems related to their functioning, physical interaction, and even ergonomics during their evaluation [44]. Additionally, using prototypes or mock-ups that do not match the aesthetic characteristics they were designed with can reduce the perceived quality and ease of use of the technology [45].

Wiedmann et al. [42] have detailed that the appearance of a product is influenced by key aspects such as color. In this sense, these authors consider that color is a factor in perceiving a concept positively or negatively. Other research has related this to visual clarity or the perception of order, alignment, and complexity in arranging of the different visual elements that make up the product. Studies focused on virtual environments, such as web pages, machine interfaces, or mobile products and applications, indicate that visual clarity promotes quick orientation in an interface and creates an impression of simplicity. It can be assumed, therefore, that this visual clarity influences usability dimensions. This is supported by Schrepp et al. [41] who found that products with better visual quality have increased efficiency and ease of learning.

However, Thielsch et al. [45] differentiate between perceived and achieved usability from product aesthetics . To further explore this, Thielsch et al. [46] conducted a study based on the results of 5 other studies. Their analysis found a small but heterogeneous influence on user performance, highlighting possible areas for future research to accurately assess this influence. Given all this, the use of one medium over another could make the cost of marketing and selling a product more profitable. However, to the best of our knowledge, there are no studies that focus on exploring the possibilities that different media can bring to industrial design and design teams for the development of successful products. This study analyzes the perception of 105 young users towards 3 models of an everyday product. The analysis is conducted through seven means of representation corresponding to three different environments (real, virtual, and conceptual). Since the selected everyday product requires a complementary item (umbrella) for its use, the study will also evaluate whether the presence of this item influences the user's perception of the products. The use or non-use of this item is considered a different setting.

III. Methodology

Through this case study, the feasibility of each medium to represent the product is analyzed to elicit an appropriate emotional response from the user. The study uses 19 parameters based on Norman's [10] three levels of emotional design (VL, BL, and RL) .

These parameters will be evaluated using seven means of product representation, which include the use of the RE, VE and CE, with and without the use of complementary items. The visualization techniques used in the study include reality, VR, VRPH, and sketches, coded as R, VR, VRPH, and S, respectively. On the other hand, when the complementary item is used, only the mediums of reality, virtual reality, and sketches are employed, represented as R+, VR+, and S+, respectively.

The results are compared qualitatively and quantitatively to identify differences and similarities between the experimental conditions and recommend the use of different visualization technique for evaluation in various design process stages.

This section has been divided in (A) a description of the case study carried out, as well as (B) the description and origin of the parameters used during the study, and (C) the sample of participating volunteers.

A. Case Study

To provide a clear understanding of the conditions under which the study was conducted, the preparation of the case study is described, including the materials (products, visualization techniques, and scenarios) used during the evaluation and the procedures followed during the user test.

1. Materials

Three different design of umbrella stands were selected as the main stimuli for the experiment. Söderman [47] concluded that prior knowledge about the product can have a negative effect into product evaluation, both in VEs and REs. For this reason, the stimuli selected was not widely consumed by the study subjects (young users). In previous investigations carried out by the some of the authors [27] with these three umbrellas stands models, it was found that the absence of the umbrella (complementary item) prevented users from recognizing the umbrella stand due, according to Galan et al. [48] to their lack of knowledge about this type of product.

As previously mentioned, the presence of the umbrella was found to impact the perception of certain characteristics of the main product, which warrants further study. To more accurately study these characteristics related to the product's usability, the use of the complementary product (the umbrella) was deemed necessary.

Additionally, previous studies [35], [36], [47], [48] have often been limited to evaluating a single product. Therefore, it was decided to conduct this research by evaluating three different models. Acording to Chu and Kao [35] set, using more than three models would have resulted in user fatigue and negatively affected the evaluation due to the extended duration of the study. In this regard, the three selected models are simple and easy to use to avoid user frustration and ensure that hand tracking is not lost during the evaluation in the virtual environment as Slater et al. state [49]. Finally, the geometry of the analyzed alternatives also helps to ensure that hand tracking is not lost during evaluation in the VE, also avoiding user frustration in this environment. On the other hand, the three products evaluated Fig. 1A had a specific cavity or position to hold long umbrellas (feature 1) and short umbrellas (feature 2). The functionalities of the selected products were similar, as well as their representation in neutral colors to avoid major differences in the user's perception. Since these products require a complementary item to fulfill their functions, two umbrellas (one large and one small) were used during experiment Fig. 1B. The umbrella stands were used in each experimental condition described above, while the complementary item was only used in R+, VR+ and S+. Physical umbrella stands were used in the REs (R and R+) and in VRPH to offer tactile feedback, while physical umbrellas were only used in R+.

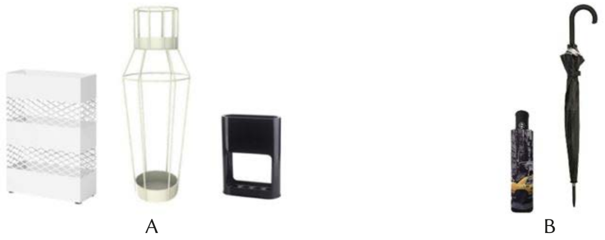


Fig. 1. Products used in the case study. Main product (A) and complementary item (B).

In the case of media representing concepts in 3D, whether physical (R and R+) or virtual (RV, RV+ and RVHP), the umbrella stand was fixed to the floor. Users could observe and touch the products, but not change their position.

The VR environment was displayed using the Oculus Quest 2 HMD, a standalone immersive VR device with a Single Fast-Switch LCD of 1832×1920 pixels per eye and a refresh rate of 72Hz. The VR environment and 3D was designed using Unity 2019.4.14f1. We used the Oculus Integration asset (version 36.0) and HPTK Posing and Snapping 2.0.0. asset for the hand tracking interaction (as the Oculus Interaction SDK was not available when the experiment was carried out). The Passthrough Capability was enabled for the calibration of the virtual objects before starting the experiment. The scene used a Real-time light with hard shadows enabled, and materials were built using a Standard Shader. The virtual objects were modelled in SolidWorks 2020, and UV mapping was completed in Blender 2.93.0.

Fig. 2A, shows the user experience in the VRPH medium. Although the volunteer perceives the umbrella stands virtually, they are synchronized with their corresponding physical models, bringing touch to the experience. Fig. 2B shows how the user has the same virtual experience, but without perceiving with touch the physical product. Finally, Fig. 2C shows how, in the same environment of the previous case, the user could manipulate the umbrellas and insert them into the different umbrella stands.

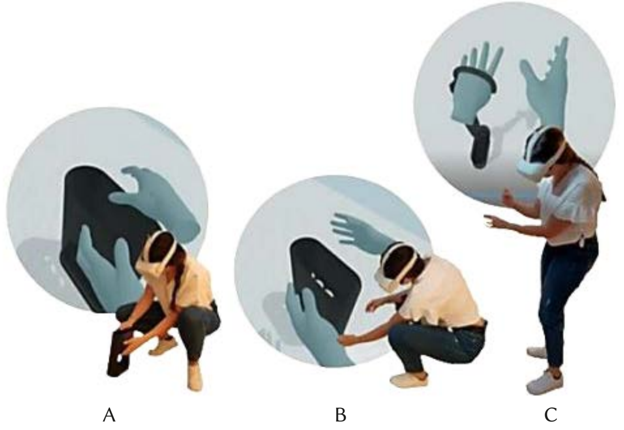


Fig. 2. User-product interaction in virtual media:VRPH(A), VR(B) and VR+(C).

As mentioned above, the user only had direct interaction with the complementary item in the R+ and VR+ media. The S+ medium, however, although it includes the representation of umbrellas, does not allow the user to interact with them. The same applies to the umbrella stands in this and S media. The static image does not allow for any interaction Fig. 3.

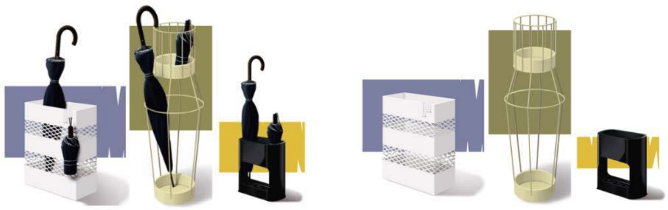


Fig. 3. Representation of products in S (left) and S+ (right) media.

In the seven visualization techniques or media described, the products were arranged in the same order and in similar spaces. For this purpose, two similar scenarios (physical rooms) of 6m2 were built, which were replicated virtually. The physical rooms were composed of 8 movable panels fixed to a 6-meter wall and are placed contiguously and symmetrically Fig. 4. The interior of the rooms was perceived by the user in a real or virtual way depending on the experimental condition in which the interaction took place. In any case, scenario 1 (red room) was built to the physical products require was built to carry out all experimental conditions that included physical products: R, R+ and VRPH. Scenario 2 (blue room) includes 3 double-sided A3 printed panels with the main products (S), and with these together with the complementary item (S+). The rest of the room was empty, to accommodate the same products and situations as the red room, but in the VEs (VR y VR+). In front of these scenarios, four seats with a table were reserved for users to comfortably fill in the questionnaires and documents required for data collection.

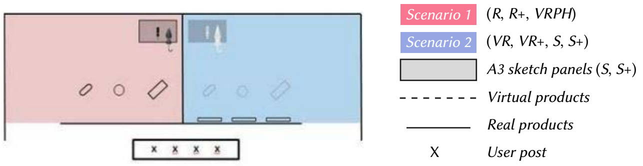


Fig. 4. Scenarios distribution.

2. Procedure

After obtaining the participant's consent and addressing any doubts or concerns with the researcher, the procedure outlined in Fig. 5 began in order to analyze 19 parameters.

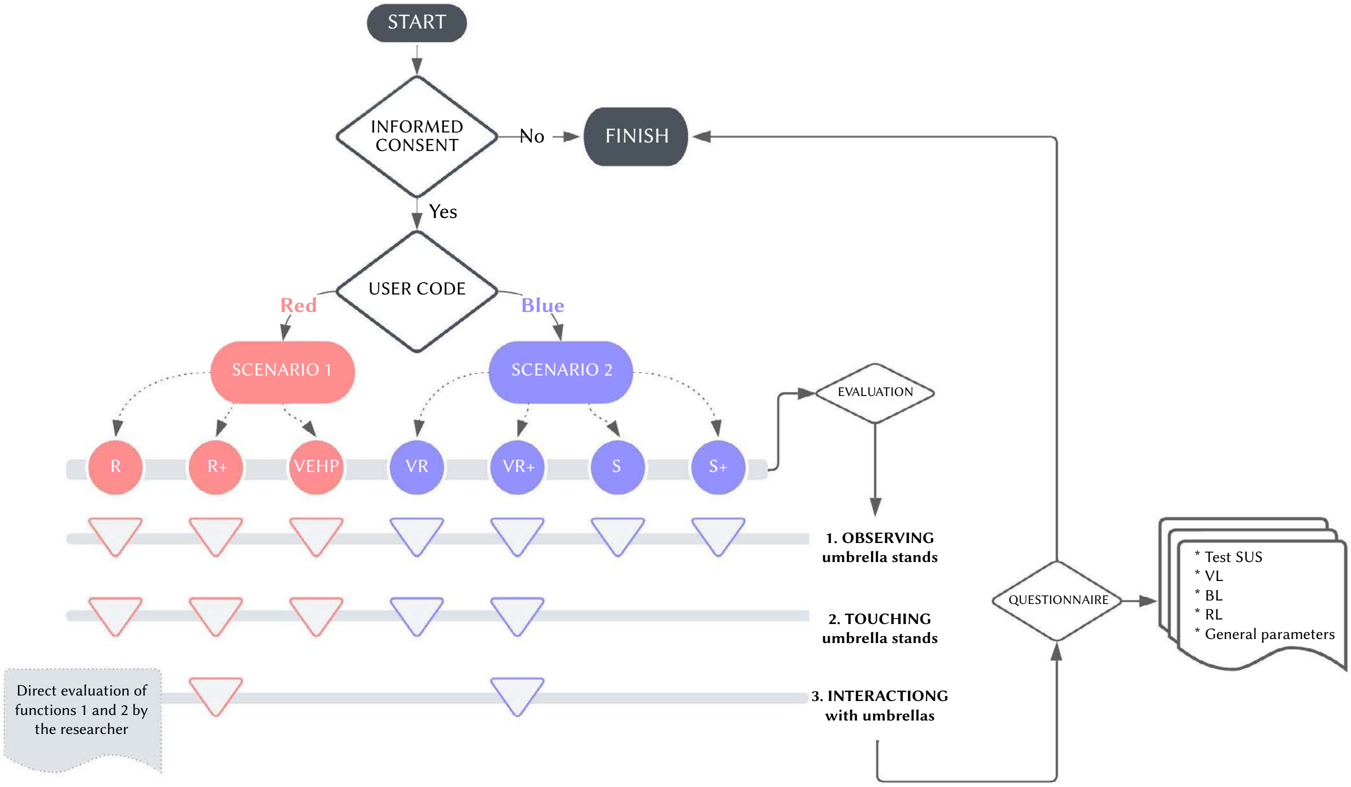


Fig. 5. Complete cycle performed by each user.

So, once the consent was signed, each volunteer received an user code with an associated colour and visualization technique. At this time, had to stand in front of the room whose color was indicated. Red corresponded to scenario 1 and blue to scenario 2.

Assisted by the researcher, each volunteer accessed the indicated space and began the evaluation process. Depending on the medium process could consist of 1, 2, or 3 phases.

In this sense, firstly, all participants first observed the three umbrella stand models. Secondly, users who enjoy the experience in the R, R+, VRPH, VR and VR+ visualization techniques proceed to touch the umbrella stands with their hands. Finally, only those who accessed through R+ and VR+ interacted with the umbrellas.

At the end of the experience, each volunteer had to fill out a questionnaire by google form. Additionally, participants who had undergone a virtual evaluation were asked to complete the Slater-Usoh-Steed presence questionnaire (SUS) [50] to quantify and qualify the level of presence in the VE. This questionnaire consisted of six 7-point Likert scale questions, with a higher score indicating a higher level of presence. It had been widely used in similar studies by other authors [27] [35] [36].

B. Evaluation Parameters

The 19 parameters established (Table I) were used for the evaluation of the selected products. These parameters (P1-P19) were used to analyze the different levels of emotional design [10]. In total, 12 parameters were used to evaluate the VL of the products (P1-P12), five for the BL (P13-P17), and two for the RL (P18-P19).

Each of these 19 parameters was evaluated in a minimum of two media (32 participants) and a maximum of seven (105 participants), based on the respective conditions. Table I indicates also the level to which they belonged, the visualization techniques (VT) used for their evaluation, method and scale used for its evaluation and the user responsible for its evaluation (data collector).

As can be seen in (Table I), to gather the users' opinion on these parameters, two different consultation methods were used. Consequently, these methods utilized two distinct scales. Firstly, P1-P14 were evaluated based on a comparison of the three different products analyzed as other researchers had done before [51], [52]. So, each participant had to establish their own "ranking" by placing the products in first, second, or third place. In this way, the results were coded on a scale from (-1) to (+1), where (-1) corresponds to the most negative value, (0) the intermediate one, and (+1) the most positive (similar to a 3 point Likert scale). Secondly, P15-P19 were evaluated individually, on a product-by-product basis. In this way, each product has been evaluated individually using a 7-point Likert scale, from (-3) to (+3), as Galan et al. [36] and Slater et al. [49].

On the other hand, to evaluate the parameters related to the user's first impression of the product (VL), the authors selected 12 bipolar pairs identifying the type of products used. For this purpose, a semantic differential was created [53] according to the procedure established by other authors [54], described in detail in subsection 1. In a comparative manner, as previously mentioned, users independently indicated their ranking positions for each product through a questionnaire. These parameters were evaluated by the users themselves following the completion of the experience, using a questionnaire (P1-P12).

These parameters have been evaluated by the user himself after the end of the experience, through a questionnaire (P1-P12).

In the specific case of the BL (P13-P17), the parameters established by Nielsen [22] and Min et al. [23] were used. These parameters, which will be described in subsections 2 and 3 of this section, were evaluated through the researcher's own observation. The researcher established P13 and P14, once again, through a ranking, the position of the different products analyzed in terms of usability. P15-P17, however, were evaluated by researchers through a 7-point Likert scale, individually.

TABLE I. Parameters Analyzed

		LEVEL	VOLUNTEERS	VT	METHOD	SCALE	DATA COLLECTOR
P1	*Light* / **Heavy**	VL	105	ALL	Ranking	3 Liker	Volunteer
P2	*Small* // Large	VL	105	ALL	Ranking	3 Liker	Volunteer
P3	*Unstable* // stable	VL	105	ALL	Ranking	3 Liker	Volunteer
P4	*Simple* / **Complex**	VL	105	ALL	Ranking	3 Liker	Volunteer
P5	*Impractical* / **practical**	VL	105	ALL	Ranking	3 Liker	Volunteer
P6	*Decorative* / **Functional**	VL	105	ALL	Ranking	3 Liker	Volunteer
P7	*Pretty* / **Ugly**	VL	105	ALL	Ranking	3 Liker	Volunteer
P8	*Modern* / **Traditional**	VL	105	ALL	Ranking	3 Liker	Volunteer
P9	*Minimalist* /**Overloaded**	VL	105	ALL	Ranking	3 Liker	Volunteer
P10	*Inexpensive* / **Expensive**	VL	105	ALL	Ranking	3 Liker	Volunteer
P11	*Vulgar* / **Elegant**	VL	105	ALL	Ranking	3 Liker	Volunteer
P12	*Common* / **Original**	VL	105	ALL	Ranking	3 Liker	Volunteer
P13	*Easy* / **difficult to learn**	BL	32	V+, R+	Ranking	3 Liker	Volunteer
P14	*Effective* / **ineffective**	BL	32	V+, R+	Ranking	3 Liker	Volunteer
P15	*Easy* / **difficult memorization**	BL	32	V+, R+	Individually	7 Liker	Researcher
P16	*Efficient* /**Inefficient**	BL	32	V+, R+	Individually	7 Liker	Researcher
P17	*Satisfactory* / **Unsatisfactory**	BL	32	V+, R+	Individually	7 Liker	Researcher
P18	Shape-assimilation	RL	105	ALL	Individually	7 Liker	Volunteer
P19	Human-assimilation	RL	105	ALL	Individually	7 Liker	Volunteer

Parameters 18 and 19 were designed to assess the recall that each product produced in the user, as well as the level of positively or negativity of the product. These parameters were also evaluated by the user himself through the questionnaire mentioned above, as P1-P12.

1. Visceral Level (P1-P12)

The VL refers to a person's first impression of a product, which is generated at a subconscious level through sensory stimuli. This initial impression cannot be controlled by the person, and the emotions evoked by the product can vary greatly. To understand the VL, three psychological processes are necessary: perception of the external world, cognition of the process of using the product, and understanding of the reflection. To measure this level, parameters are established based on the first impression that users outside of the case study have of the "umbrella stand" concept.

These parameters are based on the 12 bipolar pairs obtained from a semantic differential created using the procedure outlined in reference [54]. This was created using responses from 28 volunteers outside of the case study, including 8 professional designers with at least 5 years of experience, 12 individuals trained in industrial design, and the remaining 8 ordinary users with no design experience. Using a Google form, volunteers were asked to describe the 12 products using 5 representative adjectives. The most representative adjectives for the selected umbrella stand brands were also included, and a keyword analysis was performed. The 12 bipolar pairs are the parameters P1-P12 represented in Table I, representing the first impression that different users have of the product family to be evaluated. For the these parameters, a value closer to (-1) represented a closer correspondence with the adjective in italics, and a value closer to (+1) indicated a closer correspondence with the adjective in bold.

Based on these parameters, the first impression that each user has of the three products selected for the case study is evaluated. The user is responsible for collecting information related to these parameters through product comparison. Each volunteer ranks the three proposed models according to how well they match each parameter. Researchers should code the volunteers' responses as (-1),(0), or (+1) to conduct the data analysis. A value closer to (-1) indicated a higher match to the adjective on the left, a value closer to (+1) indicated a higher match to the adjective on the right, and (0) represents intermediate values.

2. Behavioral Level (P13-P18)

The relationship between the human being and the environment determines human behavior, which can be conscious or unconscious. In fact, in everyday life most human behavior is unconscious [55]. Although there are no specifications for its measurement, recent researchers have used the usability-related parameters [12], [16], already established by Nielsen [22]. These parameters are ease of learning (P13), which shows how easy it is to perform the tasks the first time the product is used; efficiency (P14), which evaluates the time it takes the user to perform the tasks once the product functioning is understood; memorization (P15), aimed at evaluating the errors made when performing the task; effectiveness (P16), able to recognize if after a while the user still remembers how it works; and satisfaction (P17), which seeks to know how pleasant and easy is to use the product. All these parameters are evaluated by the researcher and from the observation of the user-object interaction.

On the other hand, since these products require a complementary item for their operation, it was decided to use the latter for the correct evaluation of the main product. However, as mentioned above, it is only possible in R+ and VR+ media.

While parameters 13 and 14 were evaluated through a ranking by the comparison of the three products, parameters 15-17 were evaluated individually. For this reason, in the first case we find a 3-point liker scale, and in the second a 7-point liker scale.

As before, values with a lower score, i.e., (-1) and (-3) according to the Likert scale, had a closer correspondence with the adjective in italics. Those with a higher score, i.e., closer to (+1) and (+3) according to the Likert scale, indicated a closer correspondence with the adjective in bold.

3. Reflexive Level (P18-P19)

Emotional attachment is determined by the user's disposition to perceive, reflect on, and give meaning to a product, rather than by the product itself. According to Norman [10], this personal satisfaction in the use of a product is produced when the user experience is contrasted with previous memories, evoking an emotional response that creates a link between the user and the product. This level of attachment focuses on the user's emotions, memories, and relationships [18]. Other researchers have argued that this level of attachment can produce long-term effects related to emotions, ownership satisfaction, and the exhibition of a product [56]. To measure this level, users are asked to relate the product to objects and people in their environment based on parameters 18 and 19. Additionally, users are asked to rate the level of interest or relationship they have with these objects and people on a scale from 1 to 7.

As shown in Table I, both parameters were evaluated individually using a 7-point Likert scale. In all cases, the linkage of the product with shapes and humans which generate more negative memories in the participants has been evaluated with the score (-3). However, those products that generate the most positive memories are linked to the highest score (+3).

C. Sample

The sample was composed by engineering students between 18 and 26 years old (average of 19.1). The 64% of the sample was men, and 36% women. This fact may be due to the large glass ceiling that still exists in engineering degrees [35]. The volunteers came, as shown in Fig. 6A and Fig. 6B, from six degrees taught at the School of Industrial Engineering of the University of Málaga: Energy Engineering (EE), Industrial Technologies Engineering (ITE), Electrical Engineering (ELE), Electronic and Robotics Engineering (ERE), Industrial Design and Product Development Engineering (IDPDE) and Mechanical Engineering (ME). The participants were divided equally among the different media and situations used Fig. 6C, with the minimum sample of users participating in each medium being 14 people and the maximum being 16.

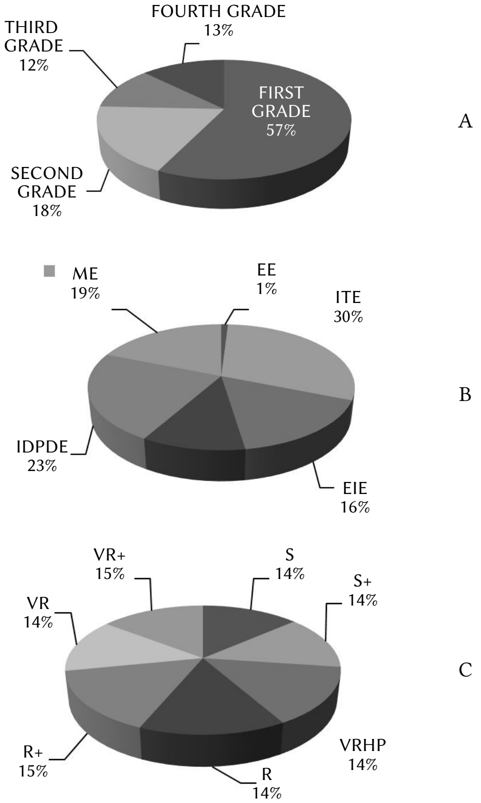


Fig. 6. Distribution of sample by grades (A), degree (B) and means of representation used (C).

The inferential analysis was performed at each level, using the parameters as dependent variables and the number of media (2 or 7) as independent variables. The same significance level (.05) and confidence interval (95%) were applied to each inferential analysis performed.

IV. Results and Discussion

To select the appropriate statistical tests, it was necessary to know the distribution of the data. The Kolmogorov-Smirnov test (for sample sizes ą50) showed that the different data sets did not follow a normal distribution, so the non-parametric Kruskal-Wallis test was used to perform an inferential statistical analysis. As mentioned before, the data has been collected through two different methods. The first consisted of a ranking comparison of the three products. Here, each user had to rank the three-stimulus studied according to the level of affinity with the evaluated parameter, as other researchers have done before [50]. The scores derived from this method were (-1), (0), or (+1) depending on whether the person ranked the product as first, second, or third, respectively, (which would be the equivalent of a 3-point Likert scale). The second was a 7-point Likert scale, where users have individually indicated how they identify the product with the analyzed bipolar pairs. This has also been used in previous studies [48]. The scores derived from this test were from (-3) to (+3).

Certain differences between the analyzed media were detected. This has been also described graphically, based on the descriptive analysis, identifying some problems in the representation of certain attributes of the products. In this context, descriptive statistics were also performed for the level of presence in the VEs. This level, measured on a 7-point Likert scale type from -3 to 3 was quite significant and similar in all the analyzed visualization techniques, although it is important to highlight that the VRPH showed the highest level of presence, being $M_{VR} = 5.18$, $M_{VR+} = 5.11$ and $M_{VRPH} = 5.32$.

A. Visceral Level

Parameters P1-P12 were analyzed through a ranking, obtaining a scale from (-1) to (+1). Fig. 7 shows the different boxplots from the descriptive analysis (one per parameter analyzed). These graphs show the distribution of the responses of the 105 participating volunteers for parameters P1-P12.

The figure suggests that there are differences between the different visualization techniques. However, it is observed that these could be due to the use of different products, and not only to the visualization technique or medium. From this, we evaluate, through an inferential analysis, the differences that may exist, by pair of visualization techniques and product to product. The evaluation difficulties of these parameters were analyzed by product using post-hoc tests with Bonferroni correction. This correction is recommended to avoid the probability of false positives in comparisons greater than 20 (Table II).

In this sense, the inferential analysis corroborates that, indeed, there were no major differences between the visualization techniques, and many of the differences seen in the figure come from the difference that the volunteers found between the products analyzed.

TABLE II. Factor P Value for Each Umbrella Stands (U1, U2, U3) and P1-P12 Parameters

PRODUCT	P1	P2	P3	P4	P5	P6
U1	p=.993	p=.528	p=.723	p=.911	p=.126	p=.495
U2	**p=.009**	**p=.013**	p=.845	p=.705	p=.731	p=.227
U3	**p=.023**	**p=.003**	p=.571	p=.348	**p=.018**	p=.233
PRODUCT	P7	P8	P9	P10	P11	P12
U1	p=.891	p=.103	p=.126	**p=.036**	p=.682	p=.712
U2	p=.479	p=.253	p=.165	p=.496	p=.155	p=.362
U3	p=.314	p=.096	p=.552	p=.170	p=.244	p=.953

Table II, where differences found are shown in bold, it is observed that only P1 (Light/heavy), P2 (Small/large), P5 (Useless/practical) and P10 (Cheap/expensive) showed differences between the visualization techniques (with a p-value < .05). In particular, differences in P1 and

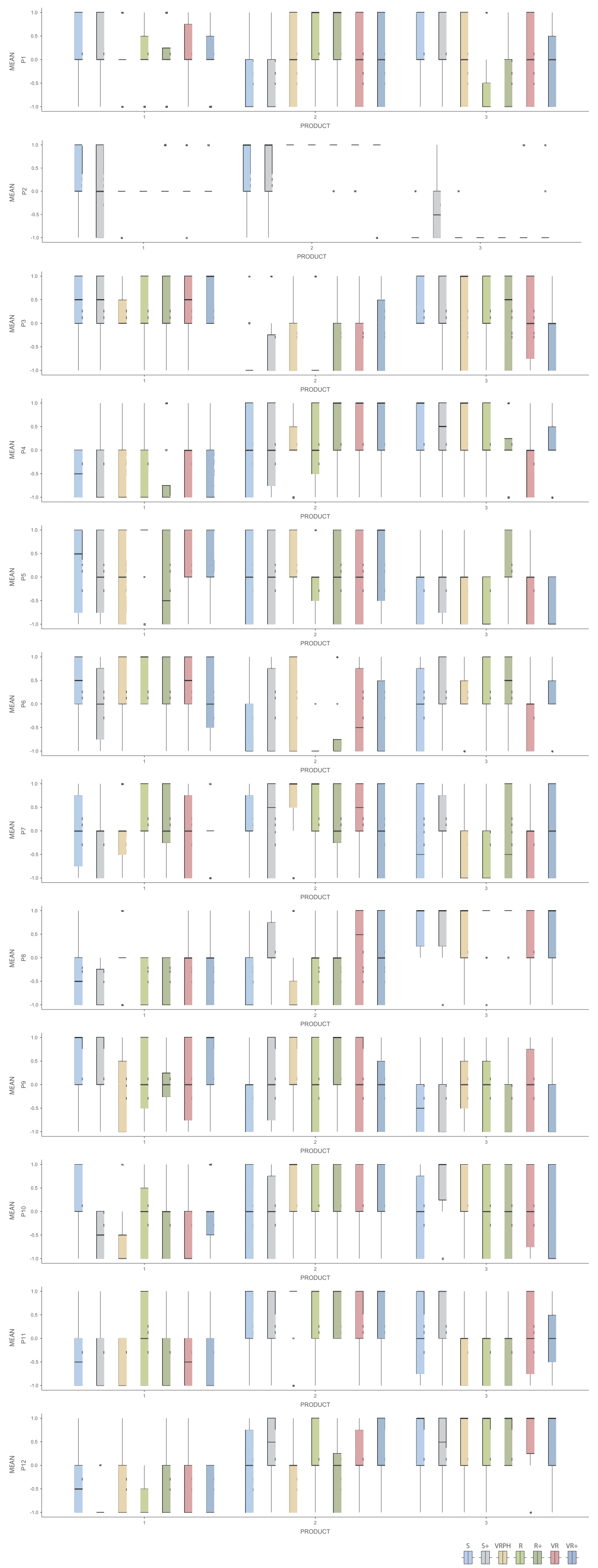


Fig. 7. Boxplot distributions of the seven visualization techniques corresponding to products 1,2 and 3 in P1-P12.

P2 are observed in Umbrella stands 2 (p-values .009 and .013) and 3 (p-values .023 and .003). However, parameters 5 and 10 only show differences in Umbrella stands 3 (p-value .018) and 1 (p-value .036), respectively.

In order to determine the media in which the greatest deviations in user perception are observed, the Post-hoc tests is performed again by visualization techniques pairs. This analysis is carried out only on those parameters that show differences (P1-P2, P5 and P10).

1. Parameters 1 and 2

Table III and Table IV show the inferential analysis for P1 and P2 respectively. In both tables, the values in the upper part correspond to umbrella stand 2, and those in the lower part to umbrella stand 3. Differences between visualization techniques are shown in **bold** by p-values.

TABLE III. Post-Hoc Tests for P1 in Products 2 (up) and 3 (Down)

P1	R	R+	VRPH	VR	RV+	S	S+
R	-	1.000	1.000	1.000	1.000	.077	**.040**
R+	1.000	-	1.000	1.000	1.000	.173	.092
VRPH	1.000	1.000	-	1.000	1.000	1.000	1.000
VR	1.000	1.000	1.000	-	1.000	1.000	1.000
VR+	1.000	1.000	1.000	1.000	-	1.000	1.000
S	.548	.718	1.000	1.000	1.000	-	1.000
S+	**.037**	.102	1.000	1.000	1.000	1.000	-

TABLE IV. Post-Hoc Tests for P2 in Products 2 (up) and 3 (Down)

P2	R	R+	VRPH	VR	RV+	S	S+
R	-	1.00	1.000	1.000	1.000	.328	**.035**
R+	1.00	-	1.000	1.000	1.000	1.000	.962
VRPH	1.000	1.000	-	1.000	1.000	.328	**.035**
VR	1.000	1.000	1.000	-	1.000	1.000	1.000
VR+	1.000	1.000	1.000	1.000	-	1.000	.710
S	1.000	1.000	1.000	1.000	.099	-	1.000
S+	**.003**	**.002**	.102	**.035**	.124	.884	-

According to this results, P1 and P2 showed differences between S+ and R. In this context, different things may be occurring with one of the visualization techniques: first, the CE may not be able to communicate the weight and dimensions of the product to the user, and secondly, the complementary item could be influencing the user's perception of these qualities, probably by confusing the viewer through the apparent resistance that the product exerts on the viewer (P1) and enlarging the volumetric space of the main stimuli through the spatial adhesion of the complementary item (P2).

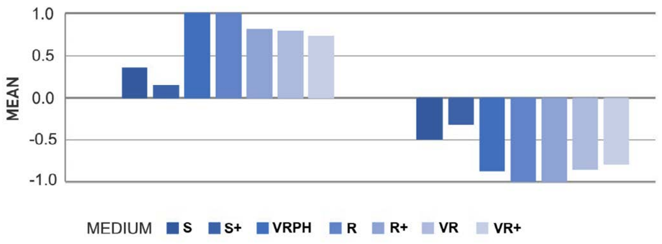


Fig. 8. Visualization techniques of parameters 1 (up) and 2 (down) by descriptive analysis in product 1 (left) and 2 (right).

Upon analyzing the results of the descriptive statistics Fig. 8, it was observed that there were no significant differences between R/R+ and S/S+. Therefore, it can be inferred that the difference detected does not depend solely on the use of the complementary item, but rather on the coordination of its use within the context of the experiment (CE). As such, it is not advisable to use S+ to measure either of the two parameters.

Additionally, even though the inferential analysis does not suggest any significant differences between S and other visual media, we do not recommend its use according to the results obtained in the descriptive analysis. Upon considering both analyses, it is recommended to use real media for P1, while for P2, either virtual or real media can be used interchangeably. This suggests that contrary to previous research [33], VR can be used professionally in the design process. However, it is important to exercise caution in its application and only use it in suitable circumstances.

2. Parameters 5 and 10

Results from P5 (Useless - Practical) and P10 (Cheap -Expensive) are shown in Table V and Table VI.

P5 showed perceptual differences for product 3 between R+/VR+ and R+/R. This seems to be due to the product itself, rather than the medium.

In R+, the product surprised the user with its ability to support the complementary items. Contradicting what may initially seem to be the case, the main product, which appeared to be light and small, could hold both large and small umbrellas. However, this capability was possible by anchoring the product to the ground. As such, all the media that were analyzed are deemed suitable for evaluating these parameters.

TABLE V. Post-Hoc Tests for P5. P Values Showing Perceptual Differences Are Shown in Red

P1	R	R+	VRPH	VR	RV+	S	S+
R	-	**.046**	1.000	1.000	1.000	1.000	.504
R+		-	1.000	1.000	**.046**	1.000	.962
VRPH			-	1.000	1.000	1.000	1.000
VR				-	1.000	1.000	1.000
VR+					-	1.000	.504
S						-	1.000
S+							-

TABLE VI. Post-Hoc Tessts for P10. P Values Showing Perceptual Differences Are Shown in Red

P2	R	R+	VRPH	VR	RV+	S	S+
R		1.000	1.000	1.000	1.000	1.000	1.000
R+		-	1.000	1.000	1.000	1.000	1.000
VRPH			-	1.000	.422	**.026**	**.048**
VR				-	1.000	.716	1.000
VR+					-	1.000	1.000
S						-	.320
S+							-

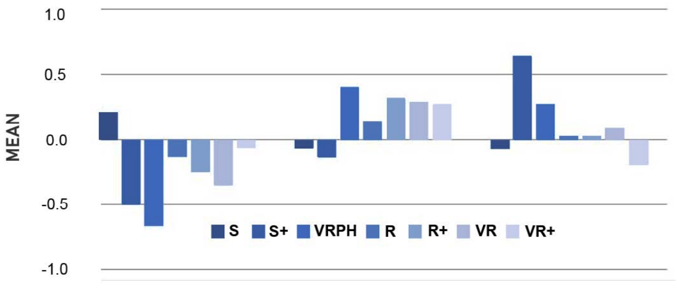


Fig. 9. Means of P10 in products 1, 2 and 3 by descriptive analysis.

The analysis showed differences in P10 for product 1 (yellow), between the S and VRPH. Comparing the results for the evaluation of this parameter between products Fig. 9, the perception of P10 in this medium was opposite between product 1 and the remaining, so that it is not recommended to consult the user about the cost of a product in S. The remaining media (including S+) are recommended as they generally produce similar values.

B. Behavioral Level

Ease of learning, efficiency, memorization, effectiveness and satisfaction (P13-P17) were evaluated based on the researcher's own observation by comparison by ranking, again obtaining a scale of -1 to 1. Additionally, P13, P16 and P17 each offer two values, according to functions 1 and 2 described above. This evaluation has only been conducted on two visual media (R+ and VR+), and the analysis is based on the experience of the 34 selected users. Therefore, the evaluation of these parameters is based on a descriptive analysis. Fig. 10 shows the different boxplots from this analysis (one per parameter analyzed).

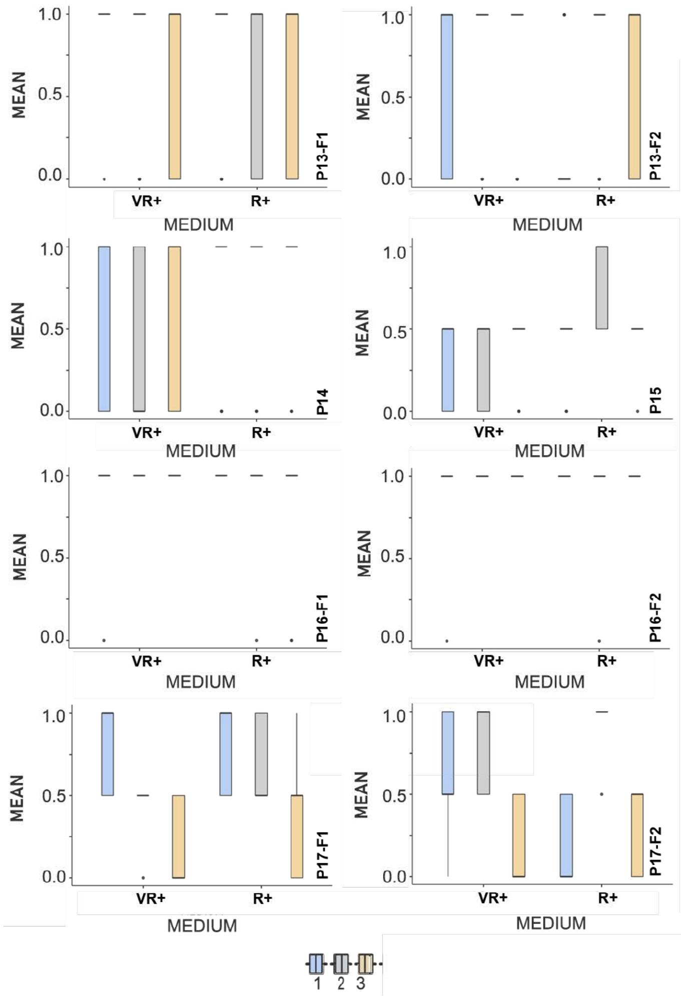


Fig. 10. Boxplot distributions of VR+ and R+ visualization techniques corresponding to products 1,2 and 3 in P3-P17.

According to this analysis, ti is possible to set that, in most of the cases, the user response was positive, except for parameters P15 and P17. Generally, P15 received negative responses in the two media analyzed, with VR+ being worse. The difference in these parameters in these two media was greater in product 2, where the user performed functions 1 and 2 without making errors. This may have been due to the need to place large umbrellas in the umbrella stands at an angle other than 90 degrees, [56] which, although more ergonomic in a real environment (RE), produced a greater number of errors in the virtual environment (VE) due to the limitation of the hand tracking technology used, as other researchers had affirmed previously [49].

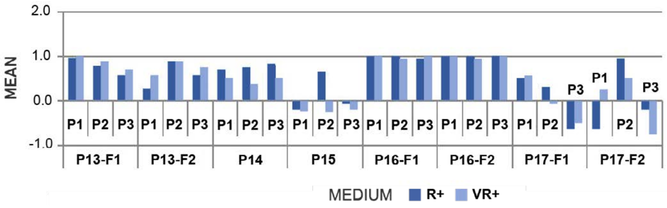


Fig. 11. Means of P13-P17 in products 1, 2 and 3 by descriptive analysis.

In the analysis of functions 1 and 2 of P17, there were significant differences between the two media used. In the case of the large umbrella, users showed greater satisfaction when using the VE, except for umbrella stand 2, which coincided with the difference discussed in P15. In the case of small umbrellas, the opposite was true. Users preferred the RE to the VE, except for umbrella stand 1, probably because the depth of the product prevented the umbrella from being positioned correctly, making it very difficult to extract the complement of the analyzed product. Therefore, in this case, we found that in product 1, the virtual evaluation was positive, while the real one was negative, contrary to what happened with the large umbrella of product 2.

Based on these data, VR+ was not recommended for the evaluation of the degree of satisfaction with the different functions of the product, nor for the detection of anthropometric errors, due to the postural differences that could be found in users who were not accustomed to the technology. However, researchers agreed with Stamps [57] that the use of VR+ could lead to significant cost and time reductions during the new product design process. This visualization technique is recommended for parameters related to functionality and usability as Liberman and Yuba set [4]. Specifically, VR could be useful for assessing ease of learning, efficiency, and effectiveness, showing insignificant differences with the other means analyzed.

C. Reflexive Level

The evaluation of shape assimilation (P18) and human assimilation (P19) was carried out using a 7-point Likert scale from (-3) to (+3). Fig. 12 shows the boxplots related two P18 and P19 parameters from the descriptive analysis. These graphs show the distribution of the responses of the 105 participating volunteers for both parameters.

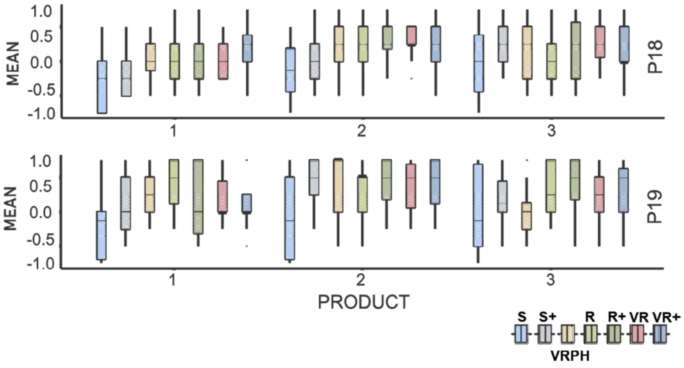


Fig. 12. Boxplot distributions of the seven visualization techniques corresponding to products 1,2 and 3 in P18-P19.

As can be seen in Table VII, no significant differences between visualization techniques were detected for any of the products evaluated, being p-values higher than .005.

TABLE VII. Factor P Value for Each Umbrella Stands (U1, U2, U3) and P19-P19 Parameters

PRODUCT	P18	P19
U1	p=.369	p=.284
U2	p=.920	p=.236
U3	p=.555	p=.274

Specifically p-values are .369, .920, and .555 for the perception of shape assimilation (P18), while .284, .236, and .274 for human assimilation (P19).

However, descriptive analysis showed that while the shape of products 1 and 3 was perceived as unpleasant across most visualization techniques, that of product 2 was perceived as desirable. Fig. 13.

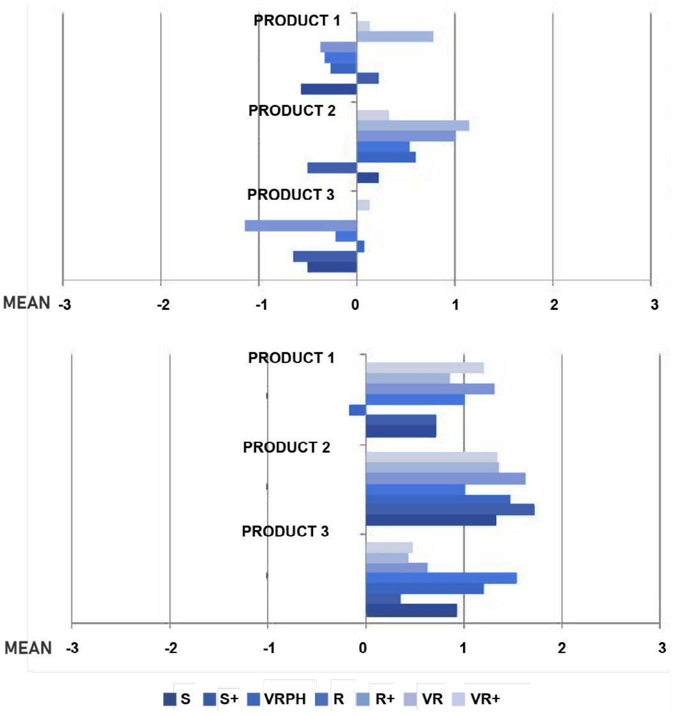


Fig. 13. Means of P18 (up) and P19 (down) in products 1, 2 and 3 by descriptive analysis.

On the other hand, regarding to P19, although users tended to relate the evaluated products with people with whom they had positive relationships, it was observed that product 1 had the lowest values, product 2 had the highest, and product 3 had intermediate values. This fact coincides with the P18 parameter, which could be because product 1 is the most traditional, and products 2 and 3 maintain a differentiating aspect with respect to other commercial umbrella stands.

V. Conclusion

Currently, the number of products on the market has been increased. This situation presents a challenge for industrial designers and practitioners, who must focus their design process on creating successful proposals. To achieve this success, it is necessary to elicit a positive emotional response from the user. To ensure this response, products must be designed and tested by users throughout the design process, from the earliest stages, and at three different levels: VL, BL and RL. However, continuous product testing can be costly in terms of time and money, and not all means of representation are suitable for evaluating user response at different levels.

This study analyzes the possibilities of seven conceptual, real, and virtual media (R, R+, VR, VR+, VRPH, S, and S+) in relation to 19 parameters to recommend to professionals the use of one or another visualization technique in the different phases of the design process.

This research has certain limitations due to the conditions of the study. First, the main products used had to be physically or virtually anchored to the ground because it was not possible to control the synchronization of the main physical product with the virtual one in the case of VRPH. In addition, all products have a drainage system that was not shown for the same reason. This prevents the user from being able to lift the objects evaluated in any of the media, which in turn precludes the perception of the "weight" parameter in the physical environment and the evaluation of the comparison of this real weight with the similar one in the virtual environment after the same action. Similarly, virtual and real environments allow the user to touch the product, while the conceptual environment does not. On the other hand, the type of products used (everyday products) may condition some aspects of the study. In addition, the usability of these everyday products can not be measured in the Virtual Reality with passive haptics visualization technique. Neither at the conceptual environment. Finally, the sample used corresponds to a specific user profile: young engineering students. Therefore, the data may not be applicable to other sectors of the population, due to the high level of familiarity of the user with new technologies.

Our results showed differences in five of them, mainly for the visceral and behavior levels. Specifically, the visualization techniques at conceptual environment (S and S+) presented difficulties for the assessment of some parameters of VL, thus these media is not recommended for the evaluation of weight. On the one hand, if the main product was represented next to the complementary item, there was a risk that the user's attention would be diverted to this item. On the other hand, if it was not applied, the user would not be able to relate the weight of the product to the resistance exerted on it, which would lead to errors in its evaluation. For this case, it is not recommended to evaluate the dimensions of a product with either medium. The sketches did not communicate the correct scale of the product without any reference (S). In the case of using the complementary item as a reference (S+), contrary to what it might seem, the differences were accentuated. The user could be confused and perceive a higher volume, perceiving the volume occupied by the products together. The sketches also indicated serious difficulties in representing the value of a product. Therefore, the use of conceptual media (S and S+) to assess its price is also discouraged.

For measuring the behavior level, R+ and VR+ are generally appropriate. Given the similarity in user response in both media, VR+ is recommended for analyzing parameters to ensure correct intuitiveness of the product, or even for training subjects due to the reduction in cost and time that may result from not making physical prototypes. However, VR+ is not recommended for evaluating the level of satisfaction with different product functions, or for detecting anthropometric errors due to postural differences in users not accustomed to the VE.

These indications are useful for researchers and companies during the design process of new products, which may also lead to the development of new work methodologies where design teams involve the user in the design process of their products.

Moreover, it would be worthwhile to further assess products utilizing the VRPH medium to mitigate the potential overestimation of perception that this medium can present across various parameters. Additionally, an intriguing analysis would be to explore multiple settings with a multifactorial analysis, selected based on the positive effects observed in this study for the most favorable settings. It should be noted that the products examined in this study were either physically or virtually anchored to the ground. In the case of VRPH, the synchronization between the primary physical product and the virtual one could not be controlled. Furthermore, all products in this study possessed a drainage system that was not displayed for the same reason.

This prevents the user from being able to lift the objects evaluated in any of the media, which in turn precludes the perception of the "weight" parameter in the physical environment and the evaluation of the comparison of this real weight with the similar one in the virtual environment after the same action. Similarly, virtual and real environments allow the user to touch the product, while the conceptual environment does not. The sample used in this study corresponds to a specific user profile: young engineering students. Therefore, the data may not be applicable to other sectors of the population, due to the high level of familiarity of the user in question with new technologies.

In addition to these limitations, it may be of interest to further explore the commercial perspective of the products. This could include studying how the presentation medium affects the user's perception of the product in the store, rather than just evaluating the product itself. This could lead to new ways of selling products through the presentation of products using different visualization technologies.

Appendix

Table VIII, Table IX and Table X provide a comprehensive analysis of the perception of three different umbrella stands based on the responses of a total of 105 volunteers. These tables present key statistical measures such as the mean, median, and standard deviation for various parameters related to the evaluation of the umbrella stands.

Table VIII focuses on parameters P1-P12 (VL) and provides insights into the perception of the umbrella stands based on the responses of 105 volunteers. It displays the average values (mean), the middle point of the dataset (median), and the measure of the spread or variability (standard deviation) for each parameter. These statistics allow us to understand the central tendency and dispersion of the responses received for each parameter.

Table IX, on the other hand, delves into parameters P13-P17 (BL), which were evaluated by a smaller subset of 35 volunteers. This table presents the mean, median, and standard deviation values for each parameter, providing a focused analysis of the perception of the umbrella stands in relation to these specific attributes.

Lastly, Table X captures the data for parameters P18-19 (RL) and summarizes the mean, median, and standard deviation values obtained from the responses of 105 participants. This table offers insights into the perception of the umbrella stands based on these parameters and allows for comparisons with the findings from the other tables.

Collectively, these tables serve as valuable tools for understanding the overall perception and variation in responses across different parameters of the umbrella stands. They provide a comprehensive analysis of the data collected, facilitating a deeper understanding of the participants' perception and preferences regarding the evaluated attributes of the umbrella stands.

TABLE VIII. Mean, Median and Standard Deviation Obtained After Descriptive Analysis of Parameters P18-P19

	VT	PRODUCT	P18	P19
Mean	B	1	3.50	4.93
		2	4.21	5.29
		3	4.00	4.71
	S+	1	3.36	4.36
		2	3.79	5.71
		3	5.07	4.71
	VRPH	1	4.07	5.20
		2	4.87	5.47
		3	4.53	3.80
	R	1	4.07	5.53
		2	4.80	5.00
		3	3.93	5.00
	R+	1	3.88	4.63
		2	5.06	5.63
		3	4.63	5.31
	VR	1	4.00	4.43
		2	5.14	5.36
		3	5.07	4.86
	VR+	1	4.67	4.47
		2	4.60	5.33
		3	4.67	5.20
Median	B	1	3.50	4.50
		2	4.50	6.00
		3	4.00	5.50
	S+	1	3.00	4.00
		2	4.00	6.00
		3	5.00	4.50
	VRPH	1	4.00	5.00
		2	5.00	7.00
		3	5.00	4.00
	R	1	4.00	6.00
		2	5.00	6.00
		3	4.00	5.00
	R+	1	4.00	4.00
		2	5.00	6.00
		3	5.00	6.00
	VR	1	4.00	4.00
		2	5.00	6.00
		3	5.00	5.00
	VR+	1	5.00	4.00
		2	5.00	6.00
		3	4.00	6.00
Standard deviation	B	1	1.51	1.44
		2	1.72	2.05
		3	2.18	2.33
	S+	1	1.34	1.60
		2	1.67	1.38
		3	1.38	1.59
	VRPH	1	1.62	1.32
		2	1.30	2.03
		3	2.10	1.70
	R	1	1.53	1.85
		2	1.42	1.81
		3	1.62	1.85
	R+	1	1.78	2.06
		2	1.12	1.59
		3	2.09	2.06
	VR	1	1.47	1.55
		2	0.949	1.69
		3	1.21	1.79
	VR+	1	1.45	1.30
		2	1.68	1.76
		3	1.63	1.70

Acknowledgment

We gratefully acknowledge the teaching staff of the Department of Graphic Expression, Design and Projects from the University of Malaga and their student volunteers for participating in this study. We also thank Paula Fernández and Javier Salgado for their help during the case study.

References

[1] Rusli, M. Aftab, "Unbroken: Rediscovering long-term value of products through change in perception," in *International Workshop on Digital Design and Manufacturing Technologies*, April 2016.

[2] R. Roy, M. Goatman, K. Khangura, "User- centric design and kansei engineering," *CIRP Journal of Manufacturing Science and Technology*, vol. 1, pp. 172–178, 2009, doi: https://doi.org/10.1016/j.cirpj.2008.10.007. Design Synthesis.

[3] P. S. Jitender, Sarkar, "Visual product assessment by using the eye-tracking equipment to study the effect of product shapes on consumer's thinking," in *Advances in Mechanical Engineering and Material Science*, April 2022, pp. 149–158, Springer Nature Singapore.

[4] E. Liberman-Pincu, Y. Bitan, "Fule—functionality, usability, look-and-feel and evaluation novel user- centered product design methodology—illustrated in the case of an autonomous medical device," *Applied Sciences*, vol. 11, p. 985, January 2021, doi: 10.3390/app11030985.

[5] J. Marquis, R. S. Deeb, "Roadmap to a successful product development," *IEEE Engineering Management Review*, vol. 46, pp. 51–58, 2018, doi: 10.1109/EMR.2018.2884275.

[6] A. Fenko, T. J. L. van Rompay, "Chapter 18 - consumer- driven product design," *Methods in Consumer Research*, vol. 2, pp. 427–462, doi: https://doi.org/10.1016/B978-0-08-101743-2.00018-2.

[7] P. W. Jordan, *Designing Pleasurable Products.* CRC Press, April 2000. M. P. Mata, S. Ahmed-Kristensen, P. B. Brockhoff,

[8] H. Yanagisawa, "Investigating the influence of product perception and geometric features," *Research in Engineering Design*, vol. 28, pp. 357–379, 2017, doi: 10.1007/s00163-016-0244-1.

[9] R. G. Cooper, "The drivers of success in new- product development," *Industrial Marketing Management*, vol. 76, pp. 36–47, 2019, doi: https://doi.org/10.1016/j.indmarman.2018.07.005.

[10] D. Norman, *Emotional Design: Why We Love or Hate Everyday Things.* Hachette, 2007.

[11] Y. Zhe, "Research on emotional design of practical ceramics," *The Journal of the Korean Society of Ceramic Art*, vol. 19, pp. 81–97, 2022.

[12] M. Aftab, H. A. Rusli, "Designing visceral, behavioural and reflective products," *Chinese Journal of Mechanical Engineering*, vol. 30, pp. 1058–1068, September 2017, doi: 10.1007/s10033-017-0161-x.

[13] E. Görnemann, S. Spiekermann, "Emotional responses to human values in technology: The case of conversational agents," *Human–Computer Interaction*, pp. 1–28, 2022, doi: 10.1080/07370024.2022.2136094.

[14] P. J. Amirkhizi, S. Pourtalebi, N. Anzabi, "Emotional effects of product form in individualist and collectivist cultures," *Journal of Marketing Communications*, pp. 1–15, February 2022, doi: 10.1080/13527266.2022.2037009.

[15] J. Yoon, A. E. Pohlmeyer, P. M. A. Desmet, C. Kim, "Designing for positive emotions: Issues and emerging research directions," *The Design Journal*, vol. 24, pp. 167–187, March 2021, doi: 10.1080/14606925.2020.1845434.

[16] N. G. Bustamante, A. A. M. Macías, A. A. Durán, J. C. O. Nicolás, A. R. Quiñones, "Usability test and cognitive analyses during the task of using wireless earphones," in *Handbook of Research on Ergonomics and Product Design*, 2018, pp. 241–263.

[17] M. Alonso-García, M. Ángel Pardo-Vicente, L. Rodríguez-Parada, D. M. Nieto, "Do products respond to user desires? a case study. errors and successes in the design process, under the umbrella of emotional design," *Symmetry*, vol. 12, p. 1350, August 2020, doi: 10.3390/sym12081350.

[18] J. Chapman, *Emotionally Durable Design.* Routledge, 2015.

[19] T. Buker, T. Schmitt, J. Miehling, S. Wartzack, "What's more important for product design – usability or emotionality? an examination of influencing factors," *Journal of Engineering Design*, vol. 33, pp. 635–669, 2022, doi: 10.1080/09544828.2022.2142902.

TABLE IX. Mean, Median and Standard Deviation Obtained After Descriptive Analysis of Parameters P1-P12 From Different Visualization Techniques (VT)

	Product	VT	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12
Mean	1	S	0.214	0.214	0.286	-0.500	0.214	0.429	0.00	-0.429	0.500	0.214	-0.357	-0.429
	2		-0.357	0.571	-0.714	0.0714	0.00	-0.357	0.0714	-0.429	-0.286	-0.0714	0.214	-0.143
	3		0.143	-0.786	0.429	0.429	-0.214	-0.0714	-0.0714	0.714	-0.357	-0.0714	0.143	0.571
	1	S+	0.143	0.00	0.286	-0.500	0.143	0.00	-0.214	-0.714	0.286	-0.500	-0.571	-0.786
	2		-0.429	0.385	-0.571	0.143	-0.0714	-0.357	0.143	0.0714	0.143	-0.143	0.143	0.357
	3		0.286	-0.357	0.286	0.357	-0.0714	0.357	0.0714	0.643	-0.429	0.643	0.429	0.429
	1	VRPH	0.0667	-0.133	0.267	-0.467	0.0667	0.133	-0.0667	-0.0667	-0.267	-0.667	-0.533	-0.467
	2		0.00	1.00	-0.400	0.133	0.267	-0.200	0.600	-0.533	0.267	0.400	0.667	-0.133
	3		0.00	-0.867	0.133	0.333	-0.333	0.0667	-0.533	0.600	0.00	0.267	-0.133	0.600
	1	R	0.0667	0.00	0.400	-0.400	0.733	0.533	0.133	-0.600	0.133	-0.133	-0.0667	-0.733
	2		0.600	1.00	-0.600	0.200	-0.0667	-0.933	0.400	-0.133	0.00	0.133	0.400	0.267
	3		-0.600	-1.00	0.200	0.200	-0.667	0.400	-0.533	0.733	-0.133	0.00	-0.267	0.467
	1	R+	0.0625	0.188	0.250	-0.563	-0.125	0.188	0.0625	-0.563	0.00	-0.250	-0.500	-0.563
	2		0.500	0.813	-0.438	0.438	0.00	-0.563	0.125	-0.375	0.500	0.313	0.688	-0.125
	3		-0.563	-1.00	0.188	0.125	0.125	0.375	-0.188	0.938	-0.438	0.00	-0.188	0.688
	1	VR	0.143	0.0714	0.429	-0.286	0.214	0.357	-0.0714	-0.357	0.0714	-0.357	-0.286	-0.571
	2		0.00	0.786	-0.571	0.429	0.00	-0.214	0.286	0.0714	0.00	0.286	0.143	0.0714
	3		-0.0714	-0.857	0.143	-0.143	-0.214	-0.143	-0.214	0.286	-0.0714	0.0714	0.143	0.571
	1	VR+	0.0667	0.0667	0.600	-0.467	0.333	0.200	-0.133	-0.400	0.467	-0.0667	-0.333	-0.667
	2		0.133	0.733	-0.333	0.400	0.333	-0.333	0.0667	-0.133	-0.133	0.267	0.400	0.200
	3		-0.200	-0.800	-0.200	0.0667	-0.667	0.133	-0.0667	0.533	-0.333	-0.200	0.00	0.467
Median	1	S	0.00	0.00	0.500	-0.500	0.500	0.500	0.00	-0.500	1.00	0.00	-0.500	-0.500
	2		-1.00	1.00	-1.00	0.00	0.00	-1.00	0.00	-1.00	0.00	0.00	0.00	0.00
	3		0.00	-1.00	0.00	1.00	0.00	0.00	-0.500	1.00	-0.500	0.00	0.00	1.00
	1	S+	0.00	0.00	0.500	-1.00	0.00	0.00	0.00	-1.00	0.00	-0.500	-1.00	-1.00
	2		-1.00	1.00	-1.00	0.00	0.00	-1.00	0.500	0.00	0.00	0.00	0.00	0.500
	3		0.00	-0.500	0.00	0.500	0.00	0.00	0.00	1.00	-1.00	1.00	1.00	0.500
	1	VRPH	0	0.00	0	-1	0	0	0	0	-1	-1	-1	-1
	2		0	1.00	-1	0	0	-1	1	-1	0	1	1	0
	3		0	-1.00	1	1	0	0	-1	1	0	0	0	1
	1	R	0	0.00	0	-1	1	1	0	-1	0	0	0	-1
	2		1	1.00	-1	0	0	-1	1	0	0	0	1	0
	3		-1	-1.00	0	0	-1	0	-1	1	0	0	0	1
	1	R+	0.00	0.00	0.00	-1.00	-0.500	0.00	0.00	-1.00	0.00	0.00	-1.00	-1.00
	2		1.00	1.00	-1.00	1.00	0.00	-1.00	0.00	0.00	1.00	0.00	1.00	0.00
	3		-1.00	-1.00	0.500	0.00	0.00	0.500	-0.500	1.00	-1.00	0.00	0.00	1.00
	1	VR	0.00	0.00	0.500	0.00	0.00	0.500	0.00	0.00	0.00	-1.00	-0.500	-1.00
	2		0.00	1.00	-1.00	1.00	0.00	-0.500	0.500	0.500	0.00	0.00	0.00	0.00
	3		0.00	-1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
	1	VR+	0	0.00	1	-1	0	0	0	0	1	0	-1	-1
	2		0	1.00	-1	1	1	-1	0	0	0	0	1	0
	3		0	-1.00	0	0	-1	0	0	1	-1	-1	0	1
Standard deviation	1	S	0.802	0.699	0.825	0.519	0.893	0.646	0.784	0.646	0.760	0.699	0.745	0.646
	2		0.842	0.646	0.611	0.917	0.877	0.842	0.730	0.756	0.726	0.917	0.802	0.864
	3		0.770	0.426	0.514	0.756	0.699	0.829	0.997	0.469	0.745	0.829	0.864	0.646
	1	S+	0.770	0.877	0.825	0.650	0.864	0.784	0.802	0.469	0.726	0.519	0.646	0.426
	2		0.852	0.768	0.756	0.864	0.917	0.929	0.949	0.730	0.864	0.864	0.770	0.745
	3		0.726	0.745	0.611	0.745	0.730	0.633	0.730	0.633	0.756	0.633	0.756	0.646
	1	VRPH	0.594	0.352	0.458	0.834	0.884	0.743	0.704	0.594	0.884	0.617	0.516	0.743
	2		0.926	0.00	0.828	0.640	0.799	1.01	0.737	0.834	0.799	0.737	0.724	0.743
	3		0.926	0.352	0.990	0.816	0.724	0.704	0.640	0.632	0.756	0.704	0.743	0.632
	1	R	0.704	0.00	0.632	0.828	0.594	0.516	0.743	0.507	0.834	0.834	0.884	0.458
	2		0.632	0.00	0.828	0.862	0.704	0.258	0.737	0.743	0.845	0.743	0.737	0.594
	3		0.737	0.00	0.676	0.676	0.488	0.632	0.743	0.594	0.834	0.926	0.799	0.834
	1	R+	0.680	0.403	0.577	0.814	0.957	0.655	0.772	0.512	0.730	0.775	0.730	0.629
	2		0.730	0.403	0.814	0.727	0.816	0.814	0.806	0.619	0.730	0.704	0.479	0.806
	3		0.727	0.00	0.911	0.619	0.719	0.719	0.911	0.250	0.727	0.894	0.750	0.479
	1	VR	0.663	0.475	0.646	0.726	0.802	0.745	0.829	0.633	0.829	0.842	0.825	0.646
	2		0.877	0.426	0.646	0.852	0.877	0.893	0.825	0.997	0.877	0.726	0.770	0.730
	3		0.917	0.535	0.864	0.770	0.802	0.770	0.802	0.726	0.829	0.829	0.864	0.756
	1	VR+	0.704	0.258	0.507	0.743	0.617	0.862	0.516	0.632	0.640	0.704	0.816	0.488
	2		0.915	0.704	0.900	0.828	0.900	0.900	0.961	0.915	0.834	0.799	0.828	0.775
	3		0.862	0.561	0.775	0.704	0.488	0.640	0.884	0.640	0.816	0.941	0.756	0.743

TABLE X. Mean, Median and Standard Deviation Obtained After Descriptive Analysis of Parameters P13-P17

	PRODUCT	MEDIUM	P13-F1	P13-F2	P14	P15	P16-F1	P16-F2	P17-F1	P17-F2
Media	1	VR+	0.941	0.588	0.647	-0.294	0.941	0.941	0.588	0.294
		R+	0.765	0.235	0.765	-0.176	1.00	1.00	0.529	-0.647
	2	VR+	0.765	0.824	0.412	-0.353	1.00	1.00	-0.0588	0.588
		R+	0.647	0.824	0.824	0.647	0.941	0.941	0.353	0.941
	3	VR+	0.647	0.824	0.647	-0.176	1.00	1.00	-0.588	-0.706
		R+	0.529	0.647	0.882	-0.0588	0.882	1.00	-0.412	-0.294
Mediana	1	VR+	1.00	1.00	1.00	0.00	1.00	1.00	1.00	0.00
		R+	1.00	0.00	1.00	0.00	1.00	1.00	1.00	-1.00
	2	VR+	1.00	1.00	0.00	0.00	1.00	1.00	0.00	1.00
		R+	1.00	1.00	1.00	1.00	1.00	1.00	0.00	1.00
	3	VR+	1.00	1.00	1.00	0.00	1.00	1.00	-1.00	-1.00
		R+	1.00	1.00	1.00	0.00	1.00	1.00	0.00	0.00
Deviation standar	1	VR+	0.243	0.507	0.493	0.470	0.243	0.243	0.507	0.588
		R+	0.437	0.437	0.437	0.393	0.00	0.00	0.514	0.493
	2	VR+	0.437	0.393	0.507	0.493	0.00	0.00	0.243	0.507
		R+	0.493	0.393	0.393	0.493	0.243	0.243	0.493	0.243
	3	VR+	0.493	0.393	0.493	0.393	0.00	0.00	0.507	0.470
		R+	0.514	0.493	0.332	0.243	0.332	0.00	0.618	0.470

[20] L. Liu, S. Cheng, H. Li, M. M. Soares, M. Li, "Usability evaluation and redesign of an integrated chair," in *HCII 2022: Design, User Experience, and Usability: UX Research, Design, and Assessment*, June 2022, pp. 428– 446.

[21] B. Boru, K. Erin, "Novel technique for control of industrial robots with wearable and contactless technologies," *Measurement*, vol. 192, p. 110850, March 2022, doi: 10.1016/j.measurement.2022.110850.

[22] J. Nielsen, *Usability Engineering*, Morgan Kaufmann, 1993, doi: 10.1016/C2009-0-21512-1.

[23] S.-H. Min, S.-W. Jeong, "Development of usability evaluation scale for manual wheelchair," *Journal of Special Education & Rehabilitation Science*, vol. 55, pp. 311–333, December 2016, doi: 10.23944/isers.2016.09.55.4.16.

[24] Y.S.Park, "The plan of the practical use of emotional design," *Korea Journal Central*, pp. 29–56, 2008.

[25] J. Ribelles, A. Lopez, V. J. Traver, "Modulating the gameplay challenge through simple visual computing elements: A cube puzzle case study," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. In Press, p. 1, 2022, doi: 10.9781/ijimai.2022.05.001.

[26] R. Hannah, S. Joshi, J. D. Summers, "A user study of interpretability of engineering design representations," *Journal of Engineering Design*, vol. 23, pp. 443–468, June 2012, doi: 10.1080/09544828.2011.615302.

[27] A. Palacios-Ibáñez, M. Alonso-García, M. Contero, J. D. Camba, "The influence of hand tracking and haptic feedback for virtual prototype evaluation in the product design process," *Journal of Mechanical Design*, vol. 145, April 2023, doi: 10.1115/1.4055952.

[28] A. T. Ranaweera, B. A. S. Martin, H. S. Jin, "What you touch, touches you: The influence of haptic attributes on consumer product impressions," *Psychology & Marketing*, vol. 38, pp. 183–195, January 2021, doi: 10.1002/mar.21433.

[29] C. D. Wood, P. K. Lewis, C. A. Mattson, "Modular product optimization to alleviate poverty: An irrigation pump case study," in *ASME 2012 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, August 2012, pp. 455–462, American Society of Mechanical Engineers.

[30] J. Vissers, D. Geerts, "Tuikit, evaluating physical and functional experiences of tangible user interface prototypes," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, April 2015, pp. 1267–1276, Association for computing Machinery.

[31] S. O. Francés, A. Villar-Aguiles, "Crazy about science, the difficulty of mixing accountability and caregiving," *Mètode Science Studies Journal*, pp. 62–73, 2018.

[32] J. Katicic, P. Häfner, J. Ovtcharova, "Methodology for emotional assessment of product design by customers in virtual reality," *Presence: Teleoperators and Virtual Environments*, vol. 24, pp. 62–73, February 2015, doi: 10.1162/PRES_a_00215.

[33] F. Górski, M. Lesik, P. Zawadzki, P. Buń, R. Wichniarek, Hamrol, "Development and studies on a virtual reality configuration tool for city bus driver workplace," in *WorldCIST 2017: Recent Advances in Information Systems and Technologies*, 2017, pp. 469–479.

[34] S. Laing, M. Apperley, "The relevance of virtual reality to communication design," *Design Studies*, vol. 71, p. 100965, November 2020, doi: 10.1016/j.destud.2020.100965.

[35] C.-H. Chu, E.-T. Kao, "A comparative study of design evaluation with virtual prototypes versus a physical product," *Applied Sciences*, vol. 10, p. 4723, July 2020, doi: 10.3390/app10144723.

[36] J. Galán, C. García-García, F. Felip, M. Contero, "Does a presentation media influence the evaluation of consumer products? a comparative study to evaluate virtual reality, virtual reality with passive haptics and a real setting," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, p. 196, 2021, doi: 10.9781/ijimai.2021.01.001.

[37] J. L. Higuera-Trujillo, J. L.-T. Maldonado, C. L. Millán, "Psychological and physiological human responses to simulated and real environments: A comparison between photographs, 360° panoramas, and virtual reality," *Applied Ergonomics*, vol. 65, pp. 398–409, November 2017, doi: 10.1016/j.apergo.2017.05.006.

[38] A. Palacios, F. Ochando, J. Camba, M. Contero, "The influence of the visualization modality on consumer perception: A case study on household products," in *13th International Conference on Applied Human Factors and Ergonomics*, 2022.

[39] L. Kent, C. Snider, J. Gopsill, B. Hicks, "Mixed reality in design prototyping: A systematic review," *Design Studies*, vol. 77, p. 101046, November 2021, doi: 10.1016/j.destud.2021.101046.

[40] M. J. M. Kamil, S. Z. Abidin, "Unconscious human behavior at visceral level of emotional design," *Procedia - Social and Behavioral Sciences*, vol. 105, pp. 149–161, 2013, doi: https://doi.org/10.1016/j.sbspro.2013.11.016.

[41] M. Schrepp, R. Otten, K. Blum, J. Thomaschewski, "What causes the dependency between perceived aesthetics and perceived usability?," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, p. 78, 2021, doi: 10.9781/ijimai.2020.12.005.

[42] K.-P. Wiedmann, J. Haase, J. Bettels, C. Reuschenbach, "It's not all about function: investigating the effects of visual appeal on the evaluation of industrial products using the example of product color," *Journal of Product & Brand Management*, vol. 28, pp. 15–27, February 2019, doi: 10.1108/JPBM-07-2017-1524.

[43] X. Zhou, P.-L. P. Rau, "Determining fidelity of mixed prototypes: Effect of media and physical interaction," *Applied Ergonomics*, vol. 80, pp. 111–118, October 2019, doi: 10.1016/j.apergo.2019.05.007.

[44] M. Galati, P. Minetola, "On the measure of the aesthetic quality of 3d printed plastic parts," *International Journal on Interactive Design and Manufacturing*, vol. 14, pp. 381– 392, June 2020, doi: 10.1007/s12008-019-00627-x.

[45] M. T. Thielsch, R. Haines, L. Flacke, "Experimental investigation on the effects of website aesthetics on user performance in different virtual tasks," *PeerJ*, vol. 7, p. e6516, February 2019, doi: 10.7717/peerj.6516.

[46] M. T. Thielsch, J. Scharfen, E. Masoudi, M. Reuter, "Visual aesthetics and performance. a first meta- analysis," in *Proceedings of Mensch und Computer 2019*, September 2019, pp. 199–210, Association for computing Machinery.

[47] M. Söderman, "Virtual reality in product evaluations with potential customers: An exploratory study comparing virtual reality with conventional product representations," *Journal of Engineering Design*, vol. 16, pp. 311–328, June 2005, doi: 10.1080/09544820500128967. doi: 10.1080/09544820500128967.

[48] J. Galán, F. Felip, C. García-García, M. Contero, "The influence of haptics when assessing household products presented in different means: a comparative study in real setting, flat display, and virtual reality environments with and without passive haptics," *Journal of Computational Design and Engineering*, vol. 8, pp. 330–342, January 2021, doi: 10.1093/jcde/qwaa081.

[49] M. Slater, A. Steed, J. McCarthy, F. Maringelli, "The influence of body movement on subjective presence in virtual environments," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 40, pp. 469–477, September 1998, doi: 10.1518/001872098779591368.

[50] C. E. Osgood, G. J. Suci, P. H. Tannenbaum, *The measurement of meaning*. University of Illinois press, 1057.

[51] Z. yong Zhou, J. ming Qi, Y. Yang, "The use of mathematical analysis in the nursing bed design evaluation," *Journal of Function Spaces*, vol. 2021, pp. 1– 10, May 2021, doi: 10.1155/2021/5520813.

[52] S. Hu, Q. Jia, L. Dong, J. Han, M. Guo, W. Guo, "An evaluation method for product design solutions for healthy aging companionship," *Frontiers in Public Health*, vol. 10, September 2022, doi: 10.3389/fpubh.2022.919300.

[53] S. H. Hsu, M. C. Chuang, C. C. Chang, "A semantic differential study of designers' and users' product form perception," *International Journal of Industrial Ergonomics*, vol. 25, pp. 375–391, May 2000, doi: 10.1016/S0169-8141(99)00026-8.

[54] M. Hua, Q. Fei, "The value of unconscious behavior on interaction design," in *10th International Conference on Computer-Aided Industrial Design & Conceptual Design*, November 2009, pp. 336–339, IEEE.

[55] B. Shneiderman, C. Plaisant, M. Cohen, N. Diakopoulos, S. Jacobs, N. Elmqvist, *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Pearson, 6 ed., 2017.

[56] A. Berni, Y. Borgianni, "Applications of virtual reality in engineering and product design: Why, what, how, when and where," *Electronics*, vol. 9, p. 1064, June 2020, doi: 10.3390/electronics9071064.

[57] A. E. Stamps, "Evaluating spaciousness in static and dynamic media," *Design Studies*, vol. 28, pp. 535–557, September 2007, doi: 10.1016/j.destud.2007.01.001.

María Alonso-García

María Alonso García is Industrial Design Engineer and Doctor from the University of Málaga, Máster of Design Engineering from National Distance Education University. She has been Professor in the Department of Mechanical Engineering and Industrial Design of University of Cádiz from 20015 to 2023 (nine coruses). Since April 2023 she is associated to the Department of Graphic Expression,Design and Projects of University of Málaga. The author focuses her research towards developing techniques and tools to support companies and design teams in the development of successful products. In this context, the researcher conducts studies to confirm the influence of multiple factors on a user's perception of an object (in controlled experiments within the design process and commercial spaces). The goal of this research is to promote a secure evaluation of concepts in the early stages of the design process that provides valuable information to companies, resulting in economic and time savings in the design process.

Almudena Palacios-Ibáñez

Almudena Palacios-Ibáñez is an Industrial Design Engineer who graduated in 2018 from the Universidad de Cádiz (UCA), Puerto Real, Spain. In 2020, she obtained a Master of Science (MSc) degree in Industrial Design from the Universitat Politècnica de València (UPV), Valencia, Spain. During this time, she engaged in early-stage research projects involving Virtual Reality (VR) in the Graphics Engineering Department to investigate its impact on brain plasticity. In the same year, she joined the UPV to pursue her PhD, supported by an FPU fellowship (FPU19/03878) by the Spanish Ministry of Education and Vocational Training. Her research project is titled "Mixed Realities and Product Perception: Experimental Analysis of the Relationship Between Visual Quality, Interaction, and User's Perceptual and Emotional Response." Furthermore, she is currently working as a Research Technician at the IUI Tecnología Centrada en el Ser Humano at UPV. Simultaneously, she is teaching Computer-Aided Design classes for the Industrial Engineering degree at UPV. Almudena's research interests primarily revolve around industrial design and product evaluation, as well as the influence of virtual reality and mixed reality technologies on product assessment.

Óscar David de-Cózar-Macías

Doctor, Industrial Organization Engineer and Industrial Technical Engineer from the University of Malaga. Associate Professor in the Department of Graphic Expression, Design and Projects. He has carried out university management tasks from 2001 to 2023 as Vice-Secretary (7 years) and Secretary (8 years) of the Higher Polytecnic School, as Director of the Secretariat of the Vice President of Smart-Campus (1,5 years) and Secretary of the School of Industrial Engineering (2,5 years), as Vice-Rector for Institutional Organization, COVID Coordinator of the University of Malaga and currently as Director of the Digital Transformation Observatory under the Vice-Rectorate for Business, Territory and Digital Transformation of the University of Malaga and the Andalusia Technology Park (PTA). In his professional role, he worked as an engineer in the company Alcatel-Citesa between 1994 and 1996. Research activity is focused on Image Analysis and Processing within the field of adjustment and recognition of geometric curves, as well as research focused on Industrial Design based on the industrial protection of products and mechanisms and the transfer of results to the business world. Among others, it is worth highlighting applied research that aims to seek transfer in the creation of patents and utility models in reference to mechanisms and industrial products obtained from an in-depth study of market needs, specifying the projects in collaboration agreements with companies. He is the principal investigator of the TEP-189 Research Group on Graphic Engineering and Design of the Junta de Andalucía and participates in Research Projects both at the regional and national level.

Manuel Damián Marín-Granados

Doctor, Industrial Organization Engineer and Industrial Technical Engineer from the University of Malaga. Associate Professor in the Department of Graphic Expression, Design and Projects. He is a highly skilled engineer and professor with a proven track record of innovation in the medical field. He has led the development of several innovative medical devices, including a goniometer for astigmatism surgery, a mechanical ventilator for COVID-19 patients, and CO2 sensors for air quality monitoring. His work has been published in leading scientific journals and has been recognized with several awards. He is the principal investigator of the TEP-189 Research Group on Graphic Engineering and Design of the Junta de Andalucía and participates in Research Projects both at the regional and national level.

User Revocation-Enabled Access Control Model Using Identity-Based Signature in the Cloud Computing Environment

Tarun Kumar[1], Prabhat Kumar[1], Suyel Namasudra[2]*

[1] Department of Computer Science and Engineering, National Institute of Technology Patna, Bihar (India)
[2] Department of Computer Science and Engineering, National Institute of Technology Agartala, Tripura (India)

* Corresponding author: suyelnamasudra@gmail.com

Received 23 December 2023 | Accepted 26 April 2024 | Published 24 May 2024

Abstract

Nowadays, a lot of data is stored in the cloud for sharing purposes across various domains. The increasing number of security issues with cloud data raises confidentiality concerns about keeping these stored or shared data. Advanced encryption and decryption techniques in cloud computing environments can be considered useful to achieve this aspect. However, an unresolved yet critical challenge in cloud data-sharing systems is the revocation of malicious users. One of the common methods for revocation involves periodically updating users' private keys. This approach increases the workload of the Key Generation Center (KGC) as the number of users increases. In this work, an efficient Revocable Identity-Based Signature (RIBS) scheme is proposed, wherein the revocation functionality is delegated to an External Revocation Server (ERS). This proposed scheme allows only the non-revoked users to access the system resources, thus, providing restricted access control. Here, the ERS generates a secret time key for signature generation based on a revoked user list. In the proposed method, a user uses its private key and secret time key to sign a message. Furthermore, to maintain data confidentiality, symmetric encryption and Elliptic Curve Cryptography (ECC) based asymmetric encryption techniques are used before outsourcing data to the cloud server. The results illustrate that the proposed scheme outperforms some of the existing schemes by providing reduced computation costs.

Keywords

Cryptography, Digital Signature, Elliptic Curve Cryptography, Identity-Based Signature, Revocation.

DOI: 10.9781/ijimai.2024.05.001

I. Introduction

In recent years, cloud computing has been offering a promising opportunity to provide computing capacity and storage for various applications. Researchers have proposed data-sharing models utilizing cloud computing across different domains. These models have facilitated the development of third-party telematics services, such as remote diagnostics, energy consumption analysis, intelligent transportation systems, and entertainment, aimed at enhancing user safety and convenience. However, before implementing these services, the crucial challenge of securing shared data in the cloud must be addressed. Unfortunately, there have been many instances, where the confidentiality of cloud data is compromised or accessed unlawfully. Due to the presence of a large number of malicious user data, cloud services must maintain strong security to prevent unauthorized access to stored data [1]. Traditional cryptographic techniques, such as RSA (Rivest Shamir Adleman) and Advanced Encryption Standard (AES), are widely used to secure data in the cloud. However, these techniques can be vulnerable to attacks like brute-force attacks. This highlights the need for robust security measures that can provide high security [2].

A digital signature is an important part of computer security that helps to identify users, verify their authenticity, and ensure they cannot deny their actions. In a typical way, the certificate authorities are responsible for managing the certificates and checking if the signature public keys are tied to the certificates. But, there is a challenge in checking if a user (certificate) has been revoked or not. In [3], Boneh and Franklin proposed a solution by regularly updating the secret keys for all non-revoked users. This approach has a critical problem, i.e., the key managing authority needs to always be online, which can create security risks and the overhead at the key managing authority dramatically increases as more users join the system. With the rise of cloud computing, many organizations are using powerful cloud servers to do heavy computing tasks. In such systems, the user revocation process is outsourced to an external server to handle the task of updating secret keys for non-revoked users.

Please cite this article as:
T. Kumar, P. Kumar, S. Namasudra. User Revocation-Enabled Access Control Model Using Identity-Based Signature in the Cloud Computing Environment, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 1, pp. 127-136, 2024, http://dx.doi.org/10.9781/ijimai.2024.05.001

Data confidentiality is another important aspect of computer security that can protect data against unauthorized access or disclosure. This can be achieved by using Deoxyribonucleic Acid (DNA) computing [4] and ECC techniques. Here, DNA sequences can be used for the secret key generation process. DNA computing is capable of computing in parallel and providing massive storage, which makes it highly efficient for handling complex, unique, and large encryption keys. ECC is also known for its efficiency and strong security properties, making it suitable for protecting sensitive data. Encrypting the data using ECC with the DNA-based generated key before transmission ensures that the data is protected during communication and the encrypted data can only be decrypted by the intended recipient having the corresponding DNA-based generated key [5]. As a result, combining DNA computing and ECC can present a novel approach for data security and privacy of sensitive information.

Enabling secure data sharing and access control along with effective user revocation management can be achieved through the integration of Identity-Based Signature (IBS) systems and ECC-based encryption techniques. These systems play a crucial role in safeguarding sensitive information and ensuring controlled access within a given framework. The integration of IBS and ECC encryption adds an extra layer of security to the data encryption process. In the context of secure data sharing, IBS mechanisms support users' authenticity and their corresponding access privileges. By utilizing user-specific identity information as the public key, these systems facilitate secure communication and data transmission processes to authorized entities. While recent works have introduced many revocable schemes, the issue of revoking identity-based signatures remains largely unexplored. Periodically updating users' private keys in identity-based approaches is typically managed by the KGC, which increases the KGC's overhead as the number of users increases in the cloud server [6].

An advanced technique has been proposed in this paper to solve the problems of the existing schemes. Here, an efficient way to handle revoked users is explored by applying identity-based signatures. In this paper, the proposed Revocable Identity-Based Signature (RIBS) scheme delegates the revocation functionality to an external revocation server. The ERS generates a secret time key that is used in the signature generation process of non-revoked users. However, if the ERS handles all the tasks of updating keys, there can be security concerns as the cloud servers are not completely trusted. So, a user's signing process is split into two parts: one uses a private key connected to their identity (given by the KGC) and the other uses a short-term secret time key connected not only to their identity, but also to the current time (given by the ERS that updates it regularly). The ERS cannot forge a digital signature because it does not have the complete signing key. To revoke a user, the KGC just informs the ERS not to issue new short-term keys. Moreover, before uploading data to the cloud server, the data owner encrypts the data using encryption techniques. In the data access phase, the data user receives the encrypted data along with a digital signature, which ensures the origin of the data. The user also receives the public key corresponding to the secret time key that is used in the signature verification process before the data encryption process. The key contributions of the proposed scheme can be summarized below.

- The proposed scheme can manage revoked users by using identity-based signatures, thus, providing access to only non-revoked or genuine users of the system.
- In this scheme, all the revocation and key generation functionalities are delegated to an external revocation server. Thus, the overhead of the central system, i.e., the cloud server, is reduced.
- Here, the DO uses a secret time key to sign the data before uploading it to the cloud server, and the corresponding public key is used to verify the signature at the user's end. This provides access control to the system as the secret time key is provided only to the authorized entities of the system.
- Additionally, the user's signing key is composed of two key components, namely the private key generated by the KGC and the short-term secret time key generated by the ERS. Therefore, neither KGC nor ERS can forge the signature as they do not have the complete signing key.

The subsequent sections of this paper are organized as follows. The Related Work section presents a comprehensive overview of existing research works. Following this, the Problem Statement section discusses the system model and key definitions of the proposed work. The methodology of the proposed work is given in detail in Section IV. Subsequently, the Performance Analysis section discusses the outcomes derived from the proposed work. Finally, the Conclusion and Future Work section summarizes the main findings of this work and outlines future research directions.

II. Related Work

The related work in data sharing, access control, and user revocation management encompasses various approaches, including identity-based signature systems, DNA encryption, and their integration for enhanced security.

Liu et al. [7] introduced a two-factor access control framework that incorporates user secret keys and security devices for accessing web-based cloud services. While this framework focuses on access control, it supports multiple layers of authentication in securing data-sharing environments. Yang et al. [8] proposed a smart card framework for multimedia cloud usage, employing a Role-Based Access Control (RBAC) model for authentication and authorization. This work provides insights into the role of smart card technologies in securing data access. Jia et al. [9] proposed a method to outsource the user revocation process to the cloud server. Instead of an immediate revocation approach, this scheme uses a periodic time-key update approach for revocation. This scheme minimizes the communication and computation costs for the key generation process. Bai et al. [10] suggested a smart card authentication framework by using ECC, wherein a single smart card serves for accessing multiple applications. By incorporating the revocation technique of [11], Tsai et al. [12] introduced the first revocable IBS scheme. In this scheme, the authors partitioned the private key of each non revoked user into two separate keys. Here, the authors verified the security strength of this scheme using standard security models. Building upon this foundation, Hung et al. [13] proposed an enhanced RIBS scheme with increased security. Sun et al. [14] later introduced an efficient RIBS scheme without pairing operations, although no formal security proof was provided by this scheme. Wei et al. [15] suggested a forward-secure RIBS scheme using the Complete Subtree (CS) method, where the KGC maintains a binary tree with each node representing a user. However, in the aforementioned RIBS schemes [12]-[15], the KGC not only issues the initial identity key for each registered user, but also periodically renews the time update keys for non-revoked users. This approach faces two challenges: (i) maintaining the KGC online poses security risks, and (ii) when the number of system users grows, the computational and communication overheads at the KGC increase rapidly. Consequently, the KGC can evolve into a security and performance bottleneck for the entire cryptosystem. Another Revocable Certificateless Public Key Encryption (RCL-PKE) scheme proposed by Ma et al. [16] addresses critical issues of the existing RCL-PKC systems. The framework of this system consists of three main entities: Private Key Generator (PKG), Cloud Revocation Agents (CRAs), and users. The PKG generates system

TABLE I. Summary of Existing Schemes

Scheme	Advantage	Disadvantage
Liu et al. [7]	This scheme provides multi-factor access control and authentication for data-sharing environments.	It cannot prevent the revoked users from accessing the system.
Tsai et al. [12]	It provides access control that solves user revocation problems.	The communication overhead of the KGC increases as the KGC is kept online always.
Hung et al. [13]	Here, a revocable access control model is proposed with improved security strength.	Both computation and communication costs increase as the number of users present in the system increases.
Yang and Lin [27]	This scheme provides confidentiality and key management by using the RSA algorithm.	The long key size used in RSA reduces the efficiency of the system in terms of storage and communication costs.
Qin et al. [28]	It provides a certificateless signing scheme and solves the key escrow issue. Here, the computation at the user end is reduced.	Pairing-based operations of this scheme result in high computation overhead. Additionally, it does not address the user revocation issue.
Liao et al. [30]	This scheme solves the issues of digital certificate management and key escrow using ECC-based operations.	It cannot revoke malicious users from the system.

parameters and distributes a secret master time key to each CRA. The PKG issues partial identity keys to users, while each CRA updates users' time update key shares based on the revocation list. Users then generate their public keys using a secret value and system parameters. By outsourcing revocation functionality to the CRA, it reduces the overhead at the PKG and introduces a dependency on the CRAs. Moreover, the use of multiple CRAs increases complexity in management and raises concerns about the overall security and reliability of the system. Yang et al. [17] introduced another secure and efficient ID-based signature scheme for the Internet of Things (IoT) environment. Though this scheme outperforms many other existing RIBS schemes in terms of computational performance, it incurs comparatively more computational cost as it is based on pairing-based operations. These pairing-based operations are very expensive computationally as compared to ECC-based scaler-point multiplication operations. Both the schemes of [16] and [17] utilize pairing-based operations for internal mathematical foundations.

In recent years, numerous schemes have been proposed by employing DNA computing to address security-related concerns, such as authentication, access control, encryption, and decryption. For instance, Adleman [18] introduced a DNA computing technique to efficiently solve the traveling salesman problem, showing the efficacy of DNA computing in solving this problem more efficiently than traditional methods. Sohal and Sharma [19] suggested a symmetric key cryptographic technique, namely Binary DNA (BDNA), that uses the concept of DNA sequences. BDNA employs a combination of DNA sequence, substitution cipher, and XOR operations to generate a secure and robust encryption key. Murugan and Thilagavathy [20] proposed a secure cloud storage model based on DNA computing and Morse code. Their approach involves storing encrypted data in a zigzag pattern to enhance data security. Addressing these limitations remains a significant challenge in the advancement of secure DNA-based computing applications [21], [22]. Apart from the above-mentioned schemes, many other schemes [23]-[30] present in the literature discuss efficient access control mechanisms for cloud computing environments. Yet, there is a need for further research to explore the potential of access control techniques for other security-related aspects in cloud computing environments [31]-[35]. Table I summarizes a few existing schemes along with their advantages and disadvantages.

III. Problem Statement

This section mainly represents the system model, design goals, and system definition of the proposed scheme in detail.

A. System Model

This work involves five different participants or entities: Key Generation Center, External Revocation Server, Data Owner (DO), Data User (DU), and Cloud Service Provider (CSP).

The key generation center plays a crucial role in setting up the system, verifying the identity of participants, and issuing key pairs. Upon an entity's request to join the system, the KGC issues the required key pair components. The external revocation server issues and updates the users' time update keys according to the Revoked User List (RUL). If a user is in the RUL, then, the ERS does not issue the secret time key for the user. A cloud service provider functions as a semi-trusted entity with significant storage and computation capabilities. It securely stores the encrypted data collected from DOs and manages access requests from data users. Data owners, including various application service providers, share data through the cloud. To ensure data confidentiality, these providers encrypt the shared data with a signature before uploading it to the cloud. The provider encrypts the data by using a DNA-based symmetric encryption algorithm as given in [36]. The provider generates a DNA-secret key and encrypts the requested data using the DNA-based symmetric encryption process. Data users can query the shared ciphertext from the cloud service provider. The workflow of the proposed model is shown in Fig. 1.

B. Design Goals

The main design goals of this work are presented below:

- **Message Confidentiality**: Data providers encrypt the shared data with a signature to ensure data confidentiality. If the user does not have a valid secret time key, which is updated periodically for each non-revoked user, then, the user cannot access the data.
- **Access Control**: By updating the secret time key and signing key of each non-revoked user, the proposed scheme restricts the revoked users from accessing the system resources.
- **Reduced Workload**: The proposed model outsources the key update process to an external server to reduce the workload of the key generation center.

C. System Definitions

The proposed scheme comprises nine algorithms, namely system initialization, registration, time key update, signature generation, data encryption, data storage phase, data access phase, signature verification, and data decryption, which are described below. The KGC maintains a RUL that contains the identities of revoked users and the RUL is updated periodically. All the symbols used in this work are described in Table II, as well as in the text.

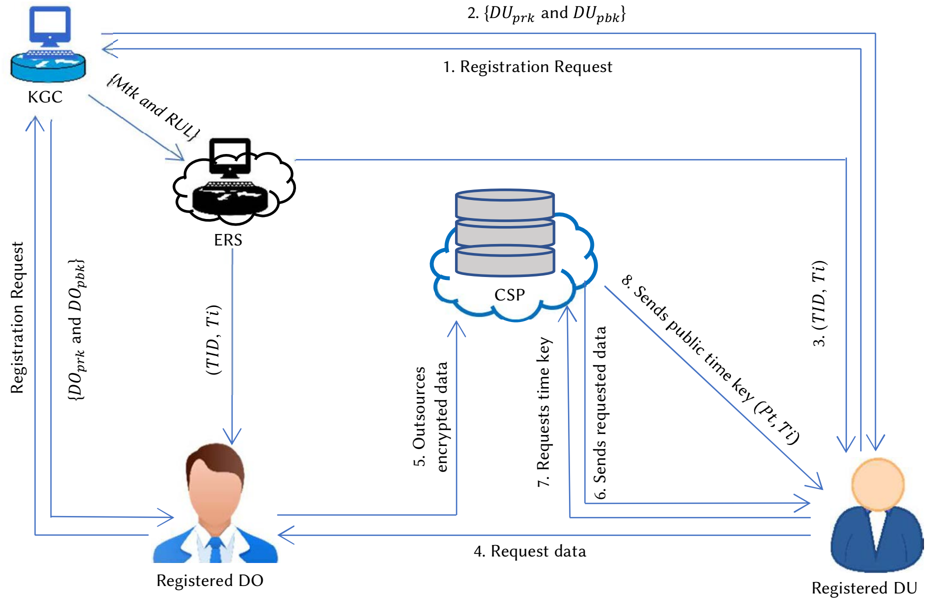


Fig. 1. System model of the proposed scheme, including all the entities.

TABLE II. Symbols and Their Description

Description	**Symbol**	**Description**	**Symbol**
KGC's Master key	KGC_{msk}	DU's private key	DU_{prk}
Secret time key at time *Ti*	(TID, Ti)	DO's private key	DO_{prk}
Hash operation	$Hash()$	DO's public key	DO_{pbk}
KGC's private key	KGC_{prk}	DU's public key	DU_{pbk}
KGC's public key	KGC_{pbk}	Signature	(m, s_1, s_2)
XOR operation	+	Hash operation	$h()$
Login Details	LD	DNA secret key	D_{SK}
DU's identity	DU_{ID}	DNA-encrypted ciphertext (CT)	CT_{SYMM}
DO's identity	DO_{ID}	Data file	PT
Master time key	mtk	Revoked and non-revoked User List	RUL, $nRUL$
Encrypted key	E_{Key}	Hash of CT	H_{CT}

System_initialization {$q \rightarrow (C(q), PP, KGC_{msk}, mtk, KGC_{prk}, KGC_{pbk})$}: The KGC performs this *System_initialization* algorithm to set up the system initially by choosing a prime number q as the security parameter. The algorithm then generates all the necessary Public Parameters (PP), including an elliptic curve $C(q)$ for the system, and the initial keys, such as KGC_{msk}, mtk, KGC_{prk} and KGC_{pbk}.

Registration {$(ID, PP) \rightarrow (prk, pbk)$}: When an entity sends a registration request, the KGC performs this *Registration* algorithm by taking entity's Identity (ID) and PP as inputs and generates private key (prk) and public key (pbk) of the entity.

Time_Key_Update:{$(PP, mtk, ID, Ti) \rightarrow (SID, Ti)$}: The ERS executes this algorithm to update the users' secret time key and to generate user's signing key (SID, Ti).

Encryption {$(D_{SK}, PT) \rightarrow CT_{SYMM}$}: The DO uses the *Encryption* algorithm to encrypt the plaintext data (PT) file and generates CT_{SYMM} by using D_{SK}.

Sign_Generation {$(PP, CT_{SYMM}, DO_{prk}, (TID, Ti)) \rightarrow (m, s_1, s_2)$}: The DO generates the signature (m, s_1, s_2) by using CT_{SYMM}, DO_{prk}, and (TID, Ti) as inputs.

Data_Storage {$(DU_{pbk}, D_{SK}) \rightarrow (E_{Key}, H_{CT})$}: The ECC encryption is used by the *DO* to encrypt D_{SK} using users' DU_{pbk}. This generates the encrypted secret key E_{Key} and H_{CT}.

Data_Access {$(REQ, DO_{ID}, DU_{ID}) \rightarrow Enc_{pt}$}: The CSP retrieves ($Pt$, Ti) from the request REQ. It encrypts (Pt, Ti) by using the users' DU_{pbk} and sends Enc_{pt} to the user.

Signature_Verification {$((m, s_1, s_2), DO_{pub}, PP) \rightarrow (Accept/Reject)$}: The user verifies the signature by using the Signature_Verification algorithm. The user takes signature (m, s_1, s_2), DO's public key (DO_{pub}), and PP as inputs and verifies the signature.

Decryption {$(D_{SK}, CT_{SYMM}) \rightarrow PT$}: The DU performs the *Decryption*() by using D_{SK} to decrypt CT_{SYMM} and generates PT.

IV. Proposed Methodology

In the beginning, the KGC initializes the system by selecting all the necessary parameters of the proposed scheme. The registration process starts with the registration requests received by the KGC, which registers all the entities, including DOs and DUs. The KGC generates login credentials, private-public key pair, and other necessary elements for all registered users. Once they are registered, the entities can request the DO to access data of their interest. If the DU is authorized, the KGC sends the requested data file to the DU along with an identity-based signature. The entire process is divided into many phases which are discussed below in detail.

A. Phase 1: System Initialization

During this phase, the central entity KGC sets up the system initially by choosing all the necessary parameters. The KGC generates its own master secret key, master time key, and private-public key pair. These keys are required to generate the necessary keys for DOs and DUs, when they register themselves. The CSP keeps its private key KGC_{msk} and *mtk* securely to maintain the system's security. The KGC sends *mtk* to the ERS. Here, *mtk* is used to generate a secret time key for each non-revoked user. This phase focuses on the initial setup and key management aspects of the cloud security system. The steps of the system setup phase executed by the KGC are as follows:

Step 1: The *KGC* selects a security parameter *q* and generates all public parameters like the elliptic curve and the generator point *P*.

Step 2: In the second step, the master key KGC_{msk} and master time key *mtk* are randomly selected by the KGC.

Step 3: Then, the *KGC* selects the private key KGC_{prk} and computes the corresponding public key $KGC_{pbk} = KGC_{prk} \cdot P$.

Step 4: Finally, the *KGC* sends *mtk* to the *ERS*.

B. Phase 2: Registration

To initiate data transmission, any entity must first register with the KGC. This involves sending a registration request to the KGC containing the entity's identity (for example, D_{ID}, i.e., *ID* of DU). Then, KGC creates a user profile and sends login details and public-private key pairs as the registration request reply. Once registered, the DU can request data from the DO and a secret time key from the ERS. By following the same procedure, a DO can also register itself with the KGC. The steps of the registration phase (for example, DU registration phase) are mentioned below.

To join the system, an entity sends a registration request to the *KGC* that includes the entity's identity (e.g., D_{ID}, representing the identity of *DU*). The KGC, in turn, creates a user profile ($DU_{ID} = h(D_{ID} + KGC_{msk})$) and inserts DU_{ID} into *nRUL* list. KGC sends public-private key pairs as a reply to the registration request. The same registration procedure applies, when the DO registers itself with the KGC.

Step 1: In the first step, the entity selects a random nonce *r*.

Step 2: Here, the entity computes $DU_{prk} = DU_{ID} + \mathrm{r} + KGC_{msk}$ and $DU_{pbk} = DU_{prk} * P$ for the data user *DU*.

Step 3: Then, it sends DU_{ID} and (DU_{prk}, DU_{pbk}) key pair to the *DU*.

Step 4: At last, the *KGC* sends the updated *RUL* and *nRUL* to the *CSP*.

By following the same procedure, the *CSP* also registers the data owners and provides DO_{prk} and DO_{pbk} key pair.

C. Phase 3: Time Key Update

Once registered, the *DU* can request data from the *DO*, as well as the secret time key from the *ERS*. Each entity (i.e., data user) needs to request for secret time keys from the ERS before starting data access. Here, to upload or store data to the cloud server, the *DO* needs to obtain its secret time key. The *DO* uses this key to encrypt the symmetric encryption key. The *DU* uses this key to verify the signature and to decrypt the encrypted symmetric key. The time key update phase as shown in Algorithm 2 provides access control to the system by providing the secret time key only to the authorized entities of the system.

Upon receiving a user's time key update request, the ERS first authenticates whether the user is a legal user by verifying its DU_{ID}, and extracts the DU's real identity D_{ID}. Before issuing the keys, the ERS again initiates the verification process by checking if the user is present in the RUL. If the user is found in the *RUL*, the ERS denies the time key request. If the user is not listed in the *RUL*, the ERS executes the algorithm and creates a new secret time key (*TID*, *Ti*) for the user. This operation involves some input parameters, namely public parameters (*PP*), the master time key (*mtk*), the user's identity (DU_{ID}), and the current time (*Ti*)). The user's signing key (*SID*, *Ti*) is composed of two components: *SID*, *Ti* = (DU_{prk} and *TID*, *Ti*). Upon receiving an update request from a user DU_{ID} at the time *Ti*, the *ERS* executes the algorithm as follows for the user:

Step 1: The ERS verifies DU_{ID} by checking if $DU_{ID} == h\,(D_{ID} + KGC_{msk})$.

Step 2: Then, it verifies if the user is a valid user by checking $DU_{ID} \in RUL$.

Step 3: In the third step, the ERS computes $h(DU_{ID})$.

Step 4: Here, the ERS computes $(TID, Ti) = mtk \cdot h\,(DU_{ID}) \cdot Ti$

```
Algorithm 1: Registration
Input: $D_{ID}$
Output: $DU_{ID}$, $RUL$, private key, and public key
Start
    If $D_{ID} \neq$ Registered entity
            Compute $DU_{ID} = h(D_{ID} + KGC_{msk})$
            Insert $DU_{ID}$ into $nRUL$
    Else
        Decline request
    If $DU_{ID} \in nRUL$
        Select $r$
        Compute $DU_{prk} = DU_{ID} + r + KGC_{msk}$
        Compute $DU_{pbk} = DU_{prk} * P$
    Else
        Invalid user
    Send $DU_{ID}$ private-public key pair to the user
    Sends updated $nRUL$ to the CSP
Stop
```

```
Algorithm 2: Generate time key
Input: $DU_{ID}$, $PP$, $mtk$, $Ti$
Output: $(SID, Ti)$
Start
    Verify $DU_{ID}$
    Compute $DU_{ID} = h(D_{ID} + KGC_{msk})$
    If $DU_{ID} == DU_{ID}$
        Check if $DU_{ID} \in RUL$
        Compute $h(DU_{ID})$
        Compute $(TID, Ti) = mtk \cdot h\,(DU_{ID}) \cdot Ti$
        Compute $Pt = (TID, Ti) * P$
        Generate $(SID, Ti) = \{DU_{prk}, (TID, Ti)\}$
        Sends $(SID, Ti)$ and $Pt$ to the user
    Else
        Invalid user
Stop
```

```
Algorithm 3: Search time key from the $CSP_{List}$
Input: $DO_{ID}$, $DU_{ID}$
Output: $Enc_{pt}$
1.  Start
2.    Verify $DU_{ID}$
3.    If $DU_{ID} \in$ Authorized user's list
4.          Goto step 7
5.    Else
6.          Stop
7.    for $i = 1$ to $U$
8.          If $DO_{ID} \in$ Authorized user's list
9.              Search $DO$ within $CSP_{List}$
10.           Retrieve DO's $(Pt, Ti)$
11.             Computes $Enc_{pt} = Enc_{ECC}(DU_{pbk}, (Pt, Ti))$
12.             Sends $Enc_{pt}$ to the $DU$
11.         Else
12.           Invalid user
13.   End for
14. Stop
```

Step 5: In the fifth step, the public key (Pt) corresponding to the secret time key (TID, Ti) is computed as $Pt = (TID, Ti) * P$.

Step 6: In the sixth step, the ERS generates the user's signing key as $(SID, Ti) = (DU_{prk}$ and $(TID, Ti))$.

Step 7: Finally, the ERS sends(SID, Ti) and Pt to the user.

D. Phase 4: Signature Generation

To sign a message CT_{SYMM}, the signer (i.e., DO) runs this algorithm with the inputs PP, CT_{SYMM}, DO_{prk}, and $\{TID, Ti\}$, and outputs the signature (m, s_1, s_2). At first, to sign a message CT_{SYMM}, the DO computes the hash of CT_{SYMM} and selects an integer k from the interval $(1, r - 1)$. The further steps of the signature generation phase (for example, DO's signature generation for message CT_{SYMM}) are as follows:

Step 1: In the first step, the DO generates a random number k.

Step 2: Here, the DO computes the hash of $CT_{SYMM} = h\ (CT_{SYMM})$.

Step 3: In this step, the DO computes $m = (k \times P)_x \bmod r$, where $(k \times P)_x$ is the x coordinate of the point $k \times P$.

Step 4: Here, $s_1 = k^{-1}\ (h\ (CT_{SYMM}) + DO_{prk} * m) \bmod r$ is computed.

Step 5: In the last step, the DO computes s_2 as

$s_2 = k^{-1}\ (h\ (CT_{SYMM}) + (TID, Ti) * m) \bmod r$, where (TID, Ti) is the DO's secret time key.

Step 6: The combined signature is (m, s_1, s_2).

E. Phase 5: Data Encryption

Here, the requested data file is encrypted by the DO by using the symmetric encryption algorithm used in [36]. The DO generates a DNA secret key D_{SK}. It encrypts the requested data PT using a symmetric encryption process and the D_{SK} to generate the ciphertext CT_{SYMM}. Then, it sends CT_{SYMM} and D_{SK} for further process, i.e., Data Storage Phase.

F. Phase 6: Data Storage

In this phase, the DO encrypts D_{SK} and generates encrypted key E_{Key} using the ECC encryption algorithm before outsourcing the complete encrypted data $\{CT_{SYMM}, E_{Key}, (m, s_1, s_2)\}\}$ to the CSP for sharing. To encrypt a message, e.g., D_{SK}, the DO selects a secret nonce n and encrypts it by using DU_{pbk}. It also includes the signature (m, s_1, s_2) corresponding to the CT_{SYMM}, which provides the authenticity and integrity of the ciphertext. The detailed steps are shown below:

Step 1: The DO generates a random nonce n.

Step 2: In the second step, the DO computes $E_{Key} = Encpt_{ECC}(DU_{pbk}, D_{SK}) = [(n \cdot P), (D_{SK} + n \cdot DU_{pbk} + Pt, Ti)]$.

Step 3: Here, the computes $H_{CT} = h(CT_{SYMM}, E_{Key})$.

Step 4: In the fourth step, CT_{SYMM}, E_{Key}, and H_{CT} are combined.

Step 5: Finally, the DO uploads $\{CT_{SYMM}, E_{Key}, H_{CT}, (m, s_1, s_2)\}$ to the CSP

G. Phase 7: Data Access

During the data access phase, the DU tries to access the received data $\{CT_{SYMM}, E_{Key}, H_{CT}, (m, s_1, s_2)\}$ from the CSP. However, to perform the signature verification and data decryption, the DU requires the DO's public key component related to its secret time key, which is embedded into s_2 component of the signature. Therefore, the DU can perform the signature verification and data decryption processes only after getting (Pt, Ti) from the CSP. When a DU sends a data access request to the CSP, the CSP verifies the DU's authenticity before providing the DO's public key component related to its secret time key (Pt, Ti) to the user. Here, the CSP searches for the DO_{ID} in the CSP's authorized user list. Then, the CSP searches the (Pt, Ti) in the CSP's list (CSP_{List}) and provides the (Pt, Ti) in the encrypted form to the requested DU. Otherwise, the CSP shows an invalid user. The detailed steps are shown below:

Step 1: At first, the DU sends a request $\{REQ, DO_{ID}, DU_{ID}\}$ to the CSP.

Step 2: The CSP verifies DU_{ID}'s authenticity.

Step 3: Here, (Pt, Ti) is retrieved by the CSP using Algorithm 3.

Step 4: In this step, the CSP computes $Enc_{pt} = Enc_{ECC}(DU_{pbk}, (Pt, Ti)$.

Step 5: The CSP sends Enc_{pt} to the DU

Step 6: Then, the DU computes $(Pt, Ti) = (Enc_{pt})$.

Step 5: Finally, the DU sends (Pt, Ti) for the sign verification process.

H. Phase 8: Signature Verification

To authenticate a signature (m, s_1, s_2) associated with the message CT_{SYMM} corresponding to the DO's identity DO_{ID} and time period Ti, the verifying entity, i.e., the DU executes this algorithm. It produces an outcome of "Accept" or "Reject" based on the validity of the provided signature. This algorithm takes signature (m, s_1, s_2), DO's public key (DO_{pub}), and PP as inputs to generate output as "Accept" or "Reject". The steps to verify a signature (m, s_1, s_2) for a message $\{CT_{SYMM}\}$ are as follows:

Step 1: In the first step, the DU computes the hash of $CT_{SYMM} = h\ (CT_{SYMM}) = e$.

Step 2: Then, m is checked by the DU regarding its validity to be an x-coordinate on the curve.

Step 3: In the third step, the DU computes $w = s_1^{-1} \bmod r$

Step 4: Then, the DU computes $u1 = (e \times w) \bmod r$ and $u2 = (r \times w) \bmod r$.

Step 5: In the fifth step, the DU computes $(x1, y1) = u1 \times P + u2 \times DO_{pbk}$.

Step 6: Then, it computes $(x2, y2) = u1 \times G + u2\ (Pt, Ti)$.

TABLE IV. Number of Cryptographic Operations Used in the Proposed Scheme

Schemes	Initial Key Extraction	Time Key Update	Sign Generation	Sign Verification
Jia et al. [9]	$T_{sm} + T_{hash}$	$T_{sm} + T_{hash}$	$2T_{sm}$	$3T_{bl} + 2T_{e}$
Tsai et al. [12]	$3T_{sm}$	$3T_{sm}$	$4T_{sm}$	$4T_{bl}$
Hung et al. [13]	$3T_{sm}$	$3T_{sm}$	$5T_{sm}$	$4T_{bl} + T_{e}$
Proposed	T_{sm}	$T_{sm} + T_{hash}$	$2T_{sm} + T_{hash}$	$2T_{sm} + T_{hash}$

Step 7: In the final step, the *DU* verifies that $m \equiv x1 \ mod\ r$ and $m \equiv x2 \ mod\ r$. If the conditions hold, the signature is valid, and the algorithm outputs "Accept".

I. Phase 9: Data Decryption

Here, the user verifies the signature using Phase 8 (Signature Verification). If the received signature matches the derived signature, the decryption process starts. At first, the user decrypts the E_{Key} by using his/her own private key. Then, the user decrypts CT_{SYMM} using the corresponding decryption process mentioned in [36]. Here, the user performs the Decryption Algorithm by using the symmetric key D_{SK} generated from the E_{Key}. The data decryption process of the proposed scheme has three steps which are represented below:

Step 1: The user first computes $DCpt\ (E_{Key}, DU_{prk})$.

Step 2: In the second step, D_{SK} is computed as $D_{SK} = Dcpt\ (Encpt_{ECC}\ (Du_{pbk}, D_{SK})) = (D_{SK} + n.\ Du_{pbk} + Pt, Ti) - (n.\ P.\ Du_{pbk} + Pt, Ti) = D_{SK}$

Step 3: Finally, the DU computes $Dcpt\ (Encpt_{ECC.}\ (D_{SK}, CT_{SYMM})) = PT$

V. Performance Analysis

This section validates the performance of the proposed scheme by giving the experimental environment and results and discussion.

A. Experimental Environment

The performance evaluation of the proposed scheme is conducted by using a cloud simulation environment developed using CloudSim 3.0.3. The experiments are executed on an HP Pro 200 G4 PC with the following specifications: Intel Core i5-10210U processor, Windows 10 Operating System, 8 GB RAM, and 2 TB HDD. Furthermore, the simulation tool, i.e., CloudSim, is configured with Apache Commons Math 3.6.1, and the HP Pro 200 G4 PC is equipped with Java version 8 for seamless compatibility and execution. Here, fifty data centers are considered to develop a cloud computing environment, which is heterogeneous. In this heterogeneous cloud environment, there are 5000 physical nodes, 4 GB memory capacity, 1 GB/s bandwidth, and four types (1. 3000 MIPS, 3 GB, 2. 4000 MIPS, 4 GB, 3. 5000 MIPS, 5 GB, and 4. 6000 MIPS, 6 GB) of VMs are considered.

B. Results and Discussion

In this subsection, a performance evaluation of the proposed scheme is presented by mainly focusing on computation cost. The proposed scheme is compared to some RIBS schemes proposed by Jia et al. [9], Tsai et al. [12], and Hung et al. [13]. The comparison among schemes is performed by considering computation costs for initial key extraction (registration and key generation in the case of the proposed scheme), time key update, signature generation, and signature verification.

Apart from these, a few other parameters like execution time for registration, time-key generation, and searching algorithms are also considered to analyze the performance of the proposed scheme. A number of experiments have been carried out to evaluate the performance of the proposed approach in comparison to other existing techniques. The experiments involve calculating the execution time for registration and searching operations for different numbers of users, ranging from 1 to 70 users as shown in Figures 3-5. Each experiment is repeated 20 times in different scenarios like for varying numbers of users and the average value is calculated to obtain accurate results. The experiments are repeated 20 times to check the stability of the results across repetitions. The performance metrics showed consistent behavior during these repetitions.

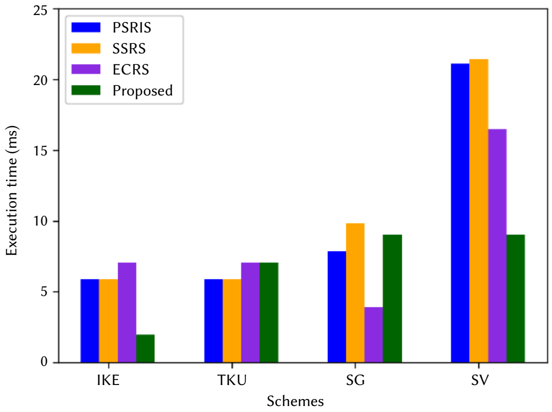


Fig. 2. Execution time (ms) in Initial Key Extraction (IKE), Time Key Update (TKU), Sign Generation (SG), and Sign Verification (SV).

To analyze the computational cost, a few operations, namely T_{sm}, T_{hash}, T_{e}, and T_{bl} as described in Table III are considered. As shown in Table IV, for the computation cost in the initial key extraction, both Tsai et al. [12] and Hung et al. [13] schemes require $3T_{sm}$(5.91ms), while Jia et al. [9] protocol requires $T_{sm} + T_{hash}$ (7.071ms). The proposed protocol requires T_{sm} (1.970ms) in this process. Regarding the time key update, Jia et al. [9] scheme requires $T_{sm} + T_{hash}$ (7.071ms), the schemes of Tsai et al. [12] and Hung et al. [13] need $3T_{sm}$ (5.91ms), while the proposed protocol requires $T_{sm} + T_{hash}$ (7.071ms). In the signing process, the scheme of Tsai et al. [28] requires $4T_{sm}$ (7.88ms), the scheme of Hung et al. [13] needs $5T_{sm}$ (9.85ms), and Jia et al. [9] scheme requires only $2T_{sm}$ (3.94ms). However, the proposed protocol requires 2 $T_{sm} + T_{hash}$ (9.041 ms), which is a little more than the schemes of Tsai et al. [28] and Jia et al. [9]. For the verification process, although Jia et al. [6] scheme involves evaluating three bilinear maps for each signature, some of them can be pre-computed. Therefore, Jia et al. [9] scheme requires $3T_{bl}+ 2T_{e}$ (16.472ms), while Tsai et al. [12] scheme requires $4T_{bl}$(21.1ms), and Hung et al. [13] scheme requires $4T_{bl} + T_{e}$ (21.431ms). Here, the proposed protocol requires $2T_{sm} + T_{hash}$ (9.041 ms) to verify the signature.

TABLE III. Execution Time of Cryptographic Operations Used in the Proposed Scheme

Operation Notation	Description	Time Cost (ms)
T_{sm}	Time cost of scalar multiplication	1.970
T_{hash}	Time cost of hash function	5.101
T_{e}	Time cost of exponential operation	0.331
T_{bl}	Time cost of bilinear pairing operation	5.270

The execution time of all phases is presented in graphical form in Fig. 2. In this figure, the schemes of Tsai et al. [12], Hung et al. [13], and Jia et al. [9] are represented by the names Provably Secure Revocable ID-Based Signature (PSRIS), Strongly Secure Revocable System (SSRS), and Efficient Cloud Revocation Server (ECRS), respectively. Here, the proposed scheme (in Fig. 2) takes comparatively slightly more time in the sign generation and verification processes. It is crucial to note that the proposed scheme generates two separate signatures, which are combined to get the actual signature. Significantly, the proposed scheme minimizes the impact on overall performance.

The execution time for registration refers to the total time required to complete the registration algorithm. Fig. 3 illustrates that the proposed approach is more efficient in terms of registration time, when compared to the existing schemes, specifically PSRIS, SSRS, and ECRS. The proposed scheme employs only one point multiplication in the entire registration and key generation processes, resulting in a less time-consuming process than the existing schemes. In contrast, PSRIS and SSRS use three point multiplication operations, while ECRS uses one hash operation and one point multiplication in the registration process. Overall, the proposed scheme supports low data encryption time compared to the existing works.

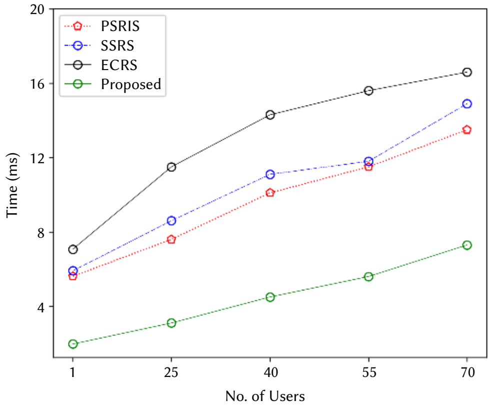


Fig. 3. Execution time for the registration process.

Fig. 4 shows the time to generate a time key for the user. In this phase, PSRIS and SSRS schemes take almost the same time to generate the time key. The proposed scheme and ECRS also take almost the same amount of time in the time key update phase, which is more than that of PSRIS and SSRS. This is because the proposed scheme uses the hash of the user's identity, i.e., DU_{ID}, to generate the time key that results in a higher execution time.

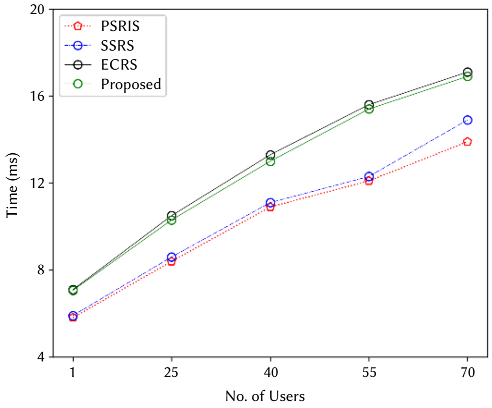


Fig. 4. Time-key generation time.

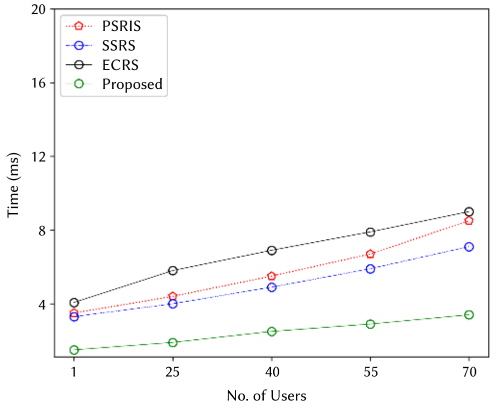


Fig. 5. Key searching time.

The next experiment is performed to compare the time for searching the secret time key (Pt, Ti) from the *CSP*. From Fig. 5, it can be seen that the proposed scheme has better performance than the other existing schemes, namely PSRIS, SSRS, and *ECRS*. In the proposed scheme, the *CSP* searches for the specified *DO*'s public key related to its time key (Pt $, Ti$) in the CSP_{List}, whenever the *CSP* receives an access request from the DU. In the CSP_{List}, (Pt, Ti) is stored along with the *DO*'s identity, i.e., DO_{ID}. Therefore, the *CSP* searches for DO's (Pt, Ti) by using his/her DO_{ID} that reduces search time in the proposed scheme. However, as the number of users with access requests increases in the system, the search time also increases, resulting in a linearly increasing graph in Fig. 5. In the existing schemes, the keys are stored and accessed by using a different approach, which takes more time as compared to the proposed scheme to search the keys. Thus, the proposed scheme facilitates faster access times, establishing its practicality for real-world applications.

VI. Conclusions and Future Work

This work has addressed critical aspects of secure data sharing, access control, and user revocation management by incorporating ECC-based techniques alongside identity-based signature systems. This scheme also uses DNA-based cryptography for symmetric key generation and encryption. The proposed scheme introduced a novel approach by delegating revocation functionality to an external revocation server, which generates a short-time secret time key that is used in the signature generation process of non-revoked users. Here, a user's signing key is generated by using two secret components: a long-term private key and a short-term secret time. This allows only authorized users to access the system resources. Additionally, the key generation center can efficiently revoke a user by just instructing the ERS not to issue a new short-term key. Moreover, in the proposed scheme, before uploading data to the cloud server, the data owner encrypts the data using ECC and symmetric encryption techniques. The experimental results show that the proposed scheme outperforms some existing schemes. Currently, this work can provide efficient access control by limiting access to only non-revoked system users. An extension of this work can be done by providing a mechanism to authenticate system entities before initiating data transmission.

References

[1] F. J. Abdullayeva, "Internet of things-based healthcare system on patient demographic data in Health 4.0," *CAAI Transactions on Intelligence Technology*, vol. 7, no. 4, pp. 644–657, 2022.

[2] L. D. Sharma, J. Rahul, A. Aggarwal, and V. K. Bohat, "An improved cardiac arrhythmia classification using stationary wavelet transform decomposed short duration QRS segment and Bi-LSTM network," *Multidimensional Systems and Signal Processing*, vol. 34, pp. 503-520, 2023.

DOI: https://doi.org/10.1007/s11045-023-00875-x.

[3] D. Boneh and M. Franklin, "Identity-based encryption from the weil pairing," in J. Kilian, (eds) *Advances in Cryptology-CRYPTO 2001. CRYPTO 2001. Lecture Notes in Computer Science*, vol. 2139. Springer, 2001. DOI: https://doi.org/10.1007/3-540-44647-8_13

[4] Y. Wang, Q. Han, G. Cui, and J. Sun, "Hiding messages based on DNA sequence and recombinant DNA technique," *IEEE Transactions on Nanotechnology*, vol. 18, pp. 299-307, 2019.

[5] L. D. Sharma, V. K. Bohat, M. Habib, A. M. Al-Zoubi, H. Faris, and I. Aljarah, "Evolutionary inspired approach for mental stress detection using EEG signal," *Expert Systems with Applications*, vol. 197, 2022. DOI: https://doi.org/10.1016/j.eswa.2022.116634

[6] Q. Qian, Y. Jia, and R. Zhang, "A lightweight RFID security protocol based on elliptic curve cryptography," *International Journal Network Security*, vol. 18, no. 2, pp. 354-361, 2016.

[7] J. K. Liu, M. H. Au, X. Huang, R. Lu and J. Li, "Fine-grained two-factor access control for web-based cloud computing services," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 3, pp. 484-497, 2016.

[8] T. C. Yang, N. W. Lo, H. T. Liaw, and W. C. Wu, "A secure smart card authentication and authorization framework using in multimedia cloud," *Multimedia Tools and Applications*, vol. 76, pp. 11715-11737, 2017.

[9] X. Jia, D. He, S. Zeadally, and L. Li, "Efficient revocable ID-based signature with cloud revocation server," *IEEE Access*, vol. 5, pp. 2945-2954, 2017.

[10] T. D. P. Bai, K. M. Raj, and S. A. Rabara, "Elliptic curve cryptography-based security framework for internet of things (IoT) enabled smart card," in *Proceedings of the World Congress on Computing and Communication Technologies* (WCCCT), IEEE, Tiruchirappalli, India, 2017.

[11] Y. M. Tseng and T. T. Tsai, "Efficient revocable id-based encryption with a public channel," *The Computer Journal*, vol. 55, no. 4, pp. 475-486, 2012.

[12] T. T. Tsai, Y. M. Tseng, and T. Y. Wu, "Provably secure revocable ID-based signature in the standard model," *Security and Communication Networks*, vol. 6, no. 10, pp. 1250-1260, 2013.

[13] Y. H. Hung, T. T. Tsai, Y. M. Tseng, and S. S. Huang, "Strongly secure revocable id-based signature without random oracles," *Information Technology and Control*, vol. 43, no. 3, pp. 264-276, 2014.

[14] Y. Sun, F. Zhang, L. Shen, and R. Deng, "Revocable identity-based signature without pairing," in *Proceedings of the* 5th *International Conference on Intelligent Networking and Collaborative Systems*, IEEE, Xi'an, China, 2013, pp. 363-365.

[15] J. Wei, W. Liu, and X. Hu, "Forward-secure identity-based signature with efficient revocation," *International Journal of Computer Mathematics*, vol. 94, no. 7, pp. 1390-1411, 2017.

[16] M. Ma, G. Shi, X. Shi, M. Su, and F. Li, "Revocable certificateless public key encryption with outsourced semi-trusted cloud revocation agent," *IEEE Access*, vol. 8, pp. 148157-148168, 2020.

[17] X. Yang, J. Wang, T. Ma, C. Chen, and C. Wang, "A secure and efficient ID-based signature scheme with revocation for IOT deployment," *in Proceedings of the Sixth International Conference on Advanced Cloud and Big Data (CBD)*, IEEE, Lanzhou, China, 2018, pp. 202-207.

[18] L. M. Adleman, "Molecular computation of solutions to combinatorial problems," *Science*, vol. 266, no. 5187, pp. 1021-1024, 1994.

[19] M. Sohal and S. Sharma, "BDNA-A DNA inspired symmetric key cryptographic technique to secure cloud computing," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 1, pp. 1417-1425, 2022.

[20] A. Murugan and R. Thilagavathy, "Cloud storage security scheme using DNA computing with morse code and zigzag pattern," *in Proceedings of the IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*, IEEE, Chennai, India, 2018, pp. 2263-2268.

[21] B. Wang, Y. Xie, S. Zhou, C. Zhou, and X. Zheng, "Reversible data hiding based on DNA computing," *Computational Intelligence and Neuroscience*, vol. 2017, 2017. DOI: http://dx.doi.org/10.1155/2017/7276084

[22] S. Namasudra and P. Roy, "Size based access control model in cloud computing," in *Proceedings of the International Conference on Electrical, Electronics, Signals, Communication and Optimization*, IEEE, Visakhapatnam, India, 2015, pp. 1-4.

[23] A. T. Ehis, "Optimization of security information and event management (SIEM) infrastructures, and events correlation/regression analysis for optimal cyber security posture," *Archives of Advanced Engineering Science*, 2023. DOI: https://doi.org/10.47852/bonviewAAES32021068

[24] V. S. Gaur, V. Sharma, and J. McAllister, "Abusive adversarial agents and attack strategies in cyber-physical systems," *CAAI Transactions on Intelligence Technology*, vol. 8, no. 1, pp. 149-165, 2023.

[25] B. M. Ahmad, S. M. Ahmed, and D. E. Sylvanus, "Enhancing phishing awareness strategy through embedded learning tools: A simulation approach," *Archives of Advanced Engineering Science*, 2023. DOI: https://doi.org/10.47852/bonviewAAES32021392

[26] D. Zhang, M. Shafiq, L. Wang, G. Srivastava, and S. Yin, "Privacy-preserving remote sensing images recognition based on limited visual cryptography," *CAAI Transactions on Intelligence Technology*, vol. 8, no. 4, pp. 1166-1177, 2023.

[27] J. H. Yang and P. Y. Lin, "A mobile payment mechanism with anonymity for cloud computing," *Journal of Systems and Software*, vol. 116, pp. 69-74, 2016.

[28] Z. Qin, J. Sun, A. Wahaballa, W. Zheng, H. Xiong, and Z. Qin, "A secure and privacy-preserving mobile wallet with outsourced verification in cloud computing," *Computer Standards & Interfaces*, vol. 54, pp. 55-60, 2017.

[29] A. J. L. Rivero, M. E. Beato, C. M. Martínez, and P. G. C. Vázquez, "Empirical analysis of ethical principles applied to different AI uses cases," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 7, pp. 105-104, 2022.

[30] Y. Liao, Y. He, F. Li, and S. Zhou, "Analysis of a mobile payment protocol with outsourced verification in cloud server and the improvement," *Computer Standards & Interfaces*, vol. 56, pp. 101-106, 2018.

[31] S. Das, S. Namasudra, S. Deb, P. M. Ger, and R. G. Crespo, "Securing IoT-based smart healthcare systems by using advanced lightweight privacy-preserving authentication scheme," *IEEE Internet of Things Journal*, vol. 10, no. 21, pp. 18486-18494, 2023.

[32] S. Khasim and S. S. Basha, "An improved fast and secure CAMEL based authenticated key in smart health care system," *Cloud Computing and Data Science*, vol. 3, no. 2, pp. 77-91, 2022.

[33] M. An, Q. Fan, H. Yu, B. An, N. Wu, H. Zhao, X. Wan, J. Li, R. Wang, J. Zhen, Q. Zou, and B. Zhao, "Blockchain technology research and application: A literature review and future trends," *Journal of Data Science and Intelligent Systems*, 2023. DOI: https://doi.org/10.47852/bonviewJDSIS32021403

[34] G. Zhang, X. Chen, L. Zhang, B. Feng, X. Guo, J. Liang, and Y. Zhang, "STAIBT: Blockchain and CP-ABE empowered secure and trusted agricultural IoT blockchain terminal," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 5, pp. 66-75, 2022.

[35] G. Thakur, P. Kumar, C. M. Chen, A. V. Vasilakos, Anchna, and S. Prajapat, "A robust privacy-preserving ECC-based three-factor authentication scheme for metaverse environment," *Computer Communications*, vol. 211, pp. 271-285, 2023.

[36] T. Kumar, S. Namasudra, and P. Kumar, "Providing data security using DNA computing in the cloud computing environment," *International Journal of Web and Grid Services*, vol. 19, no. 4, pp. 463-486, 2023.

Tarun Kumar

Tarun Kumar is a research scholar in the Department of Computer Science and Engineering at the National Institute of Technology Patna, Bihar, India. He is also an Assistant Professor in the School of Computing Science and Engineering, Galgotias University, Greater Noida, India. He has 16 years of experience in academics. His research interests are cloud computing and DNA computing. He has published several papers in peer reviewed journals and international conferences. He also organized and attended several workshops.

Prabhat Kumar

Prabhat Kumar is a Professor in the Computer Science and Engineering Department at the National Institute of Technology Patna, India. He is a senior member of IEEE, a Professional member of ACM, a life member of CSI, IAENG, ISTE, and a global member of the internet society. He has more than 130 publications in reputed international journals, conference proceedings, and book chapters. He has supervised six PhD scholars, 25 M. Tech., scholars, and has an Indian patent on his name. His research area includes wireless sensor networks, internet of things, data analytics, software engineering, e-governance, and many more.

Suyel Namasudra

Suyel Namasudra has received Ph.D. degree from the National Institute of Technology Silchar, Assam, India. He was a post-doctorate fellow at the International University of La Rioja (UNIR), Spain. Currently, Dr. Namasudra is working as an assistant professor in the Department of Computer Science and Engineering at the National Institute of Technology Agartala, Tripura, India. Before joining the National Institute of Technology Agartala, Dr. Namasudra was an assistant professor in the Department of Computer Science and Engineering at the National Institute of Technology Patna, Bihar, India. His research interests include blockchain technology, cloud computing, IoT, AI, and DNA computing. Dr. Namasudra has edited 6 books, 5 patents, and 80 publications in conference proceedings, book chapters, and refereed journals like IEEE TCE, IEEE TII, IEEE T-ITS, IEEE TSC, IEEE TCSS, IEEE TCBB, ACM TOMM, ACM TOSN, ACM TALLIP, FGCS, CAEE, and many more. He is the Editor-in-Chief of the Cloud Computing and Data Science (ISSN: 2737-4092 (online)) journal. Dr. Namasudra has served as a Lead Guest Editor/Guest Editor in many reputed journals like IEEE TBD (IEEE, IF: 7.2), ACM TOMM (ACM, IF: 3.144), MONET (Springer, IF: 3.426), CAEE (Elsevier, IF: 3.818), CAIS (Springer, IF: 4.927), CMC (Tech Science Press, IF: 3.772), Sensors (MDPI, IF: 3.576), and many more. He has also participated in many international conferences as an organizer and session chair. Dr. Namasudra is a member of IEEE, ACM, and IEI. He has been featured in the list of the top 2% scientists in the world in 2021, 2022, and 2023. His h-index is 36.

Machine Learning for Financial Prediction Under Regime Change Using Technical Analysis: A Systematic Review

Andrés L. Suárez-Cetrulo[1], David Quintana[2], Alejandro Cervantes[3]*

[1] Ireland's Centre for Applied AI (CeADAR), University College Dublin (Ireland)
[2] Department of Computer Science and Engineering, Universidad Carlos III de Madrid, Avda. Universidad 30, 28911 Leganes (Spain)
[3] Escuela Superior de Ingeniería y Tecnología, Universidad Internacional de La Rioja (UNIR), Logroño (Spain)

* Corresponding author: alejandro.cervantesrovira@unir.net

Received 26 May 2022 | Accepted 28 April 2023 | Published 23 June 2023

unir
LA UNIVERSIDAD
EN INTERNET

Abstract

Recent crises, recessions and bubbles have stressed the non-stationary nature and the presence of drastic structural changes in the financial domain. The most recent literature suggests the use of conventional machine learning and statistical approaches in this context. Unfortunately, several of these techniques are unable or slow to adapt to changes in the price-generation process. This study aims to survey the relevant literature on Machine Learning for financial prediction under regime change employing a systematic approach. It reviews key papers with a special emphasis on technical analysis. The study discusses the growing number of contributions that are bridging the gap between two separate communities, one focused on data stream learning and the other on economic research. However, it also makes apparent that we are still in an early stage . The range of machine learning algorithms that have been tested in this domain is very wide, but the results of the study do not suggest that currently there is a specific technique that is clearly dominant.

Keywords

Concept Drift, Finance, Machine Learning, Meta Learning, Regime Change, Systematic Literature Review.

DOI: 10.9781/ijimai.2023.06.003

I. Introduction

Financial markets can be described as an evolutionary and nonlinear dynamical complex system [1], [2]. Forecasting in the financial domain has traditionally been performed under the assumption that the underlying data has been created by a linear process [3]. Another line of work to make financial predictions is to use machine learning (ML). These algorithms have surprised financial experts [4]–[6] because of their success in mapping nonlinear relationships without prior knowledge [7]. Deep learning algorithms (neural networks) and ensembles have been some of the techniques obtaining the best results for stock trend prediction [8]–[13].

Different crises, recessions and bubbles, such as the COVID-19 pandemic, or volatile mid-term trends in crypto markets, have made apparent the non-stationary nature and the presence of drastic structural changes in financial markets [14]. During these periods, mean returns, volatility and correlations among assets tend to change quickly [15]. This has brought attention to the problem of concept drift [16] in computational finance [17]. Many recent research works point out that financial assets or companies present different states that may repeat or not overtime or evolve due to inflation, deflation, or changes in supply and demand [18]–[24].

In finance, a change in the collective behaviour of market participants and their reactions is called a regime change (RC). As covered by the marked efficiency hypothesis [25], we cannot observe the individual behaviour of a trader or its intentions. Instead, we can only observe changes in the price dynamics and macro or micro-economic variables and extrapolate the changes that make them modify their behaviour. The execution of these strategies is the actual generative process of the observed time series of prices or trends. The estimation of the hidden processes driving the market into different regimes is often approached using regime-switching models, a type of time series model where parameters can have different values in different cycles [26].

Despite the fact that artificial intelligence has recently become a trend and even a buzzword in many industries, this has not become the main trend yet for trading systems. This is mainly due to the high complexity and hard explainability of these models [27], being the second a must for stakeholders and decision-makers in this sector [28]. Instead, traders tend to identify directional changes in the market state using different popular indicators tailored according to their needs.

Traditionally, the literature has used static methods to interpret patterns based on the meaning of these indicators and their historical correlation to future prices. However, this correlation may vary over

Please cite this article as:
A. L. Suárez-Cetrulo, D. Quintana, A. Cervantes. Machine Learning for Financial Prediction Under Regime Change Using Technical Analysis: A Systematic Review, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 1, pp. 137-148, 2024, http://dx.doi.org/10.9781/ijimai.2023.06.003

time. Behavioural shifts of investors changing continuously with a hidden context can also be observed through the change in sell versus buy volumes, in differences between local minima and local maxima over time, and through different moving averages at different time frames depending on the granular detail observed (frequencies) at intraday, daily or weekly levels. Changes in financial markets challenge traders and investors, as most of their models rely on previous patterns. Hence, a way to recognise these changes provides a competitive advantage since it allows changes in trading strategies ahead of other investors [29]. Detecting concept shifts also helps lower the risks of financial exposure in high-frequency trading (HFT).

The digitalisation of the financial industry has resulted in a growing amount of data that is available for decision-making. This, together with the increasing amount of computational resources, has accelerated the adoption of a whole range of machine-learning-based solutions. Among the suite of instruments available to deal with regime changes, online incremental ML algorithms seem especially appropriate. Among the advantages that they offer, we can mention the fact they can handle non-stationarities, shifts, and drifts in price generation processes. Another aspect that makes them a good fit for this context is the fact that they are scalable for continuous learning scenarios [30].

One might consider two main scenarios regarding the nature of structural change. The first possibility is the existence of recurrence, that is, the idea that the system might transition back to a previous price generation process. For instance, there might be a specific market state for market openings at intraday frequencies and another for financial bubbles that might be observable at lower frequencies. The alternative assumes that any drift results in a transition to a new process. As we will discuss in detail, while there is a relevant number of published studies on machine learning for data streams that pay attention to non-stationarity [31]–[35], the literature on financial applications of these algorithms, especially that focused on recurring concept drifts, is more limited [17], [36]–[41].

We must also point out that the prediction of future financial trends can be tackled using fundamental or technical analysis. Despite some controversy regarding its potential [25], [42], the latter is very prevalent in short-term trading [43], hence the focus on this approach. Having said that, there are also relevant papers in the first category, like the contributions of Geva and Zahavi [44], on the short-term, intraday and high-frequency forecast using news data and the study of Dogra et al. [4], analysing the impact of recent news on stock price trends and challenges such as class imbalance. More recently, Chen et al. [45] hybridised both approaches in a study that combines both sentiment analysis and technical indicators.

With this survey, we try to bring to the academic community's attention how ML is being used to deal with structural change in financial markets. Our goal is to identify directions on leveraging the benefits of modern algorithms that work with different scenarios and deal with any changes that may arise in real-time.

The use of these approaches may help to find strategies to improve prediction accuracy during times of change, limiting the need for constant model retraining. Hence, some of these techniques efficiently increase the potential profits, avoiding the computational burden and benefiting from always having an up-to-date model.

The rest of the document is structured as follows: Section II covers the methodology and research questions used in this systematic review; Section IV will discuss the outcomes of each of the research questions. There, Subsection A will introduce the topic of regime change in financial series, and Section B will discuss the core literature on machine learning for financial prediction under regime change. The final two sections will be reserved for a summary of results, main conclusions, and future research lines.

II. Research Methodology

A. Motivation and Objectives

Financial time series are often subject to structural change. Even though machine learning offers major advantages in this context, the literature on the topic is limited and sparse. There seem to be different research communities focused on different aspects of the problem, and it is hard to keep track of the main contributions and instruments used to tackle the problem.

The literature presents a lack of studies on prediction under regime changes based on technical analysis using machine learning. This is unfortunate, as there is a lot to be gained in terms of efficiency and performance. Within this field, we find very promising ideas. For instance, the problem can be framed using the data stream learning topic of concept drift. A significant number of contributions to this new concept have not been explicitly applied to finance yet, and they are not widely known. They have not been widely present in machine learning surveys outside the data stream learning niche area.

Exploring previous research showed that a comprehensive review does not exist on these topics. Therefore, this study will help readers understand the current state of the art, bridge the gap among research fields, and address promising future lines of research in this domain.

B. Research Method

In order to provide an overview of the state of the question, our research has followed Kitchenham and Charters' guidelines on Systematic Literature Review (SLR) [46], [47].

A systematic review is defined as an organised way to synthesise existing work fairly. An SLR is a means to identify, evaluate and interpret the available research works relevant to a definite research question, topic area, or phenomenon of interest. After revising the literature for similar research objectives, it can be identified that there is no previously published search on a topic.

C. Planning

The study aims to summarise the current status of predicting financial time series in the financial literature during behavioural or regime changes in markets. Kitchenham and Charters' SLR protocol was adapted to describe the plan for the review.

The protocol comprises research background and questions, search strategy, study selection criteria and procedures, data extraction, and data synthesis strategies to guarantee that the investigation is undertaken as intended and reduce the likelihood of bias in the study. In this protocol, the entire investigation plan was not decided from the beginning. Instead, this and the results produced were recorded as the study progressed.

D. Research Questions

This paper has the following two research questions:

Q1 What are the different research areas for predicting under regime changes in the financial literature? and

Q2 What are the most commonly used machine learning techniques applied to analysing regime changes?

The results expected at the end of the systematic review were to see what research or surveys had been applied or produced on the topic so far and to identify the implications of using machine learning to handle behavioural changes in financial markets in the scientific literature.

E. Search Strategy and Process

The search strategy included: i) search resources and ii) a search process. Each one of them is detailed in the following subsections.

1. Search Resources

This study was planned to find all the literature available about machine learning for forecasting under regime changes in finance. The sources used for the systematic review were:

- IEEE Digital Library (http://ieeexplore.ieee.org);
- ScienceDirect, on the subject of Computer Science (https://www.sciencedirect.com/);
- ACM Digital Library (http://portal.acm.org);
- Taylor & Francis Journals (http://www.tandfonline.com);
- Wiley Online Library (http://www.wiley.com/);
- SpringerLink (http://link.springer.com); and additionally
- Google Scholar was explored as grey literature (https: //scholar.google.com/).

2. Search Process

The overall search process is depicted in Fig. 1 and is explained in the following section.

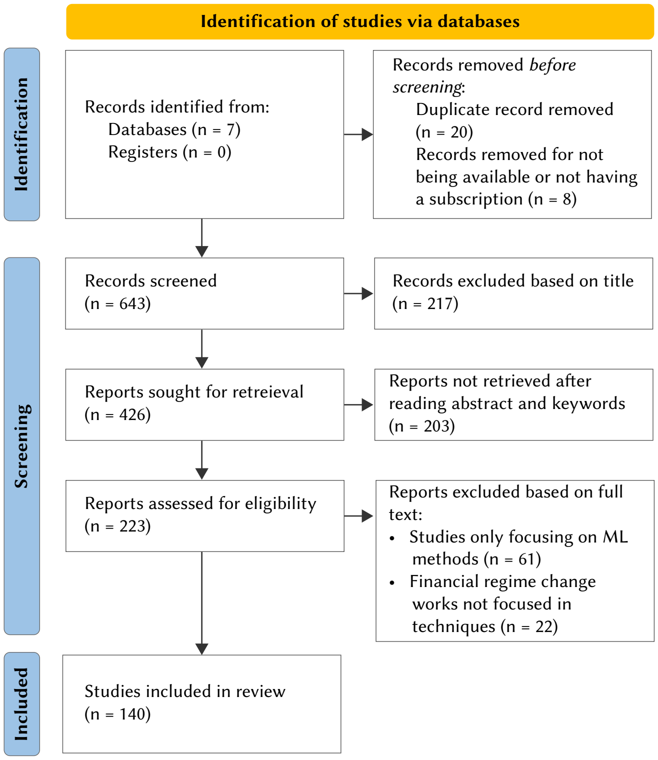

Identification of studies via databases
Identification
Records identified from:
Databases (n = 7)
Registers (n = 0)
Records removed *before screening*:
Duplicate record removed (n = 20)
Records removed for not being available or not having a subscription (n = 8)
Screening
Records screened (n = 643)
Records excluded based on title (n = 217)
Reports sought for retreieval (n = 426)
Reports not retrieved after reading abstract and keywords (n = 203)
Reports assessed for eligibility (n = 223)
Reports excluded based on full text:
• Studies only focusing on ML methods (n = 61)
• Financial regime change works not focused in techniques (n = 22)
Included
Studies included in review (n = 140)

Fig. 1. Flow of information through the different phases of the review using a PRISMA diagram [48].

The starting point was choosing a set of relevant keywords. They were: *regime change, regime-switching model, machine learning, change detection* and *stock trend forecasting*. The search was then run on the already mentioned databases in March 2022, returning 643 works in total in a time range, including the years 1970 to 2022. Irrelevant and duplicate publications were removed, and 223 unique research works remained. At that point, publications were reviewed based on titles, abstracts, conclusions, references and keywords and then were classified into three different types:

- Relevant works: these should satisfy one of the two inclusion criteria covered later in this subsection;
- Process assessment works: if the publication is related to the financial domain or concept drift literature and is relevant.
- Excluded works: works not relevant to the topic.

When there was doubt about the classification of a research workpiece, it was included in the relevant group, leaving the possibility of discarding it during the next stage, when the full-text versions were reviewed. Third, each full article was retrieved and read to verify its inclusion or exclusion. The reason for exclusion or inclusion in this third stage was documented. Fourth, to check the consistency of the inclusion/exclusion decisions, a test-retest approach and re-evaluation of a random sample of the primary studies were made.

Documents were kept when they satisfied at least one of the criteria below:

- The work was explicitly related to regime changes or structural breaks in non-stationary data.
- The work was relevant to machine learning forecasting in domains with complex dynamics and non-stationarities in the financial field.

The authors reviewed all 223 research works and put them into these different groups according to the previously mentioned criteria. This list was reviewed to check for inconsistencies. The result of this stage was that 140 publications were classified as relevant.

There is a risk that some relevant works have been missed. Therefore, this study cannot guarantee completeness. However, it can still be trusted to give a good overview of the relevant literature on price forecasting in the financial domain under structural breaks.

3. Data Extraction

The data extracted from each publication was documented and kept in a reference manager. After the identification of the publications, the following was extracted:

- Source (journal, book, conference or strictly relevant technical or white paper);
- Title;
- Publication year;
- Authors;
- Classification according to topics;
- Summary of the research, including which questions were solved.

III. Summary of Results

In order to analyse the 223 works, we found the need to classify them in more ways than just according to the methodology defined in Section II. When needed, the topics were updated or clarified during the classification process. Results of the classification process with regard to the research questions are detailed in Table I.

TABLE I. Classification of Papers With Regard to the Research Questions

Question	Topic	Relevant Studies	Quantity
Q1	Regime changes	[14], [15], [18], [19], [23], [26], [29], [49]–[63]	22
Q1 and Q2	ML in stock forecasting	[1]–[13], [20]–[22], [24], [25], [27], [28], [37], [38], [42]–[45], [64]–[110]	73
Q2	Concept drift and online ML	[16], [17], [30]–[36], [39]–[41], [111]–[143]	45

The data required for analysis were extracted by exploring the full text of each research work. Table II presents the results of the search and the source of the documents. Table III presents the results in the second stage. As mentioned before, the total number of papers remaining after the exclusion process was 140. Table I summarises their classification according to the knowledge area.

The relevance of regime changes or structural breaks in the literature of financial price forecasting leads to consider two major areas: financial regime changes (related to majorly statistical approaches to detect change points or forecast under different regimes) and data stream learning (where the problem of concept drift can be understood as a type of regime change in the ML literature).

TABLE II. Results Without Filtering

Data Source	Total Publications
ScienceDirect	76
Google Scholar	56
Springerlink	33
IEEE Digital Library	27
ACM Digital Library	14
Wiley	9
Taylor & Francis	8

TABLE III. Second Stage Results

Data Source	Total Publications
ScienceDirect	56
Google Scholar	34
Springerlink	20
IEEE Digital Library	16
Taylor & Francis	5
Wiley	5
ACM Digital Library	4

Fig. 2 shows how, out of a total of 140 relevant studies, the majority of the works reviewed to correspond to ML techniques applied to stock forecasting. Some of these works overlap regime change research, focusing primarily on probabilistic models to classify directional changes and represent different regimes. The literature on online learning does not tend to coincide with the one on regime changes. However, studies of online ML tackle similar challenges as models to handle regime changes, such as having up-to-date models and re-training mechanisms. A deeper discussion on this matter will be held in Section IV.

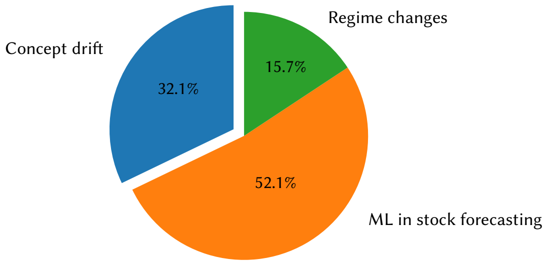


Fig. 2. Topic distribution of research papers after filtering.

Fig. 3 shows the distribution of papers reviewed across various sources. A majority of the research works have been retrieved from high-impact journals, followed by conferences and books. However, since some of the topics reviewed, like online ML, are current research areas, a remainder, close to « 4% of research works, belong to non-peer-reviewed papers contained by open-access repositories.

Finally, we have extracted from the papers classified papers under the topic "Concept Drift" the ML technique mainly used, either as a new proposal or as a reference for comparison. We have grouped these techniques into eight broad categories (Fig. 4 and 5). For this task, we have excluded reviews. These results show that most of the reviewed papers use techniques from four main categories: Evolving systems (that include Evolving clustering, Evolving fuzzy rules and Fuzzy neuro systems), Ensemble based systems (usually with tree-based components), traditional systems adapted to concept change (such as adaptive decision trees), and finally Neural Networks and Deep Learning. The latter are more recent in general, and therefore this trend is likely to become more important in the near future. Fig. 6 shows the evolution of these categories.

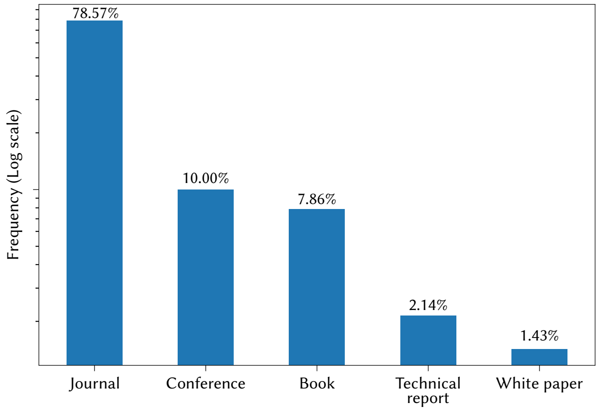


Fig. 3. Source distribution of research papers after filtering.

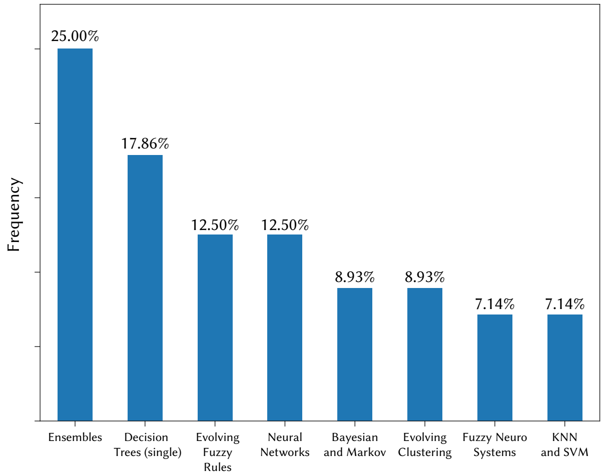


Fig. 4. ML techniques found in Concept Drift studies, grouped by categories, counting each different technique used in papers and assigning the corresponding category separately. In this figure, a single paper comparing several algorithms in the same category is counted as many times as algorithms.

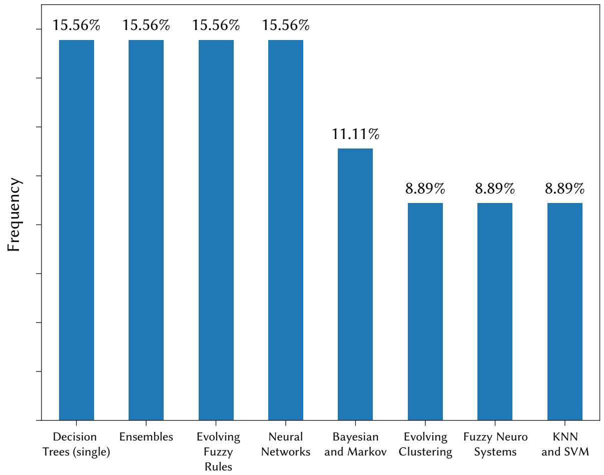


Fig. 5. Types of ML found in Concept Drift studies, using the unique categories found in the same paper. In this figure, a single paper comparing six methods in category A and one method in category B is counted as only two entries (one for A and another for B).

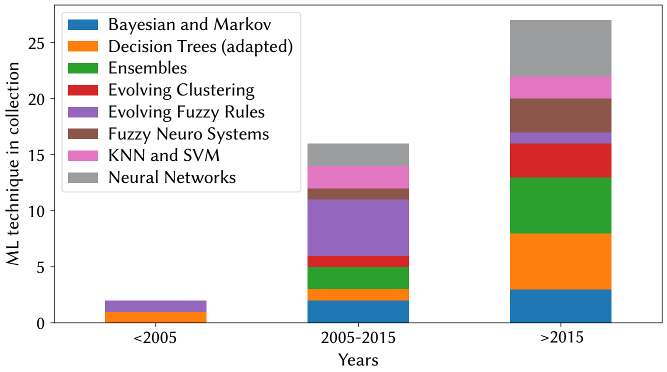


Fig. 6. Categories per period of 10 years based on data used in Fig. 5.

IV. Discussion

This section describes the papers reviewed in this work. In this discussion, we follow the schema in Fig. 7.

A. Regime Changes in Financial Series (Q1)

Early studies from the financial literature claim that financial markets are efficient [25] and, as a result, asset prices follow a random walk [81]. Fama [25] claimed that markets cannot be consistently beaten on a risk-adjusted basis and that their prices cannot be anticipated has always been a source of controversy in the literature. Many research works have pointed to different markets being predictable using different sources of information [5], [77], [78], [85], [88], [104].

Forecasting in the financial domain can be characterised by a non-stationary and unstructured nature and by hidden relationships [2], [74]. Economic, social and political factors within countries and international impact add uncertainty to financial markets [66], [70], [79], [93], [94], [100], [106]. Hence, markets can be considered an evolutionary and nonlinear complex system [1]. The financial literature has covered different approaches to predicting market prices using statistical and, more recently, AI-based methods.

In recent years, different events like the COVID-19 pandemic or the bankruptcy of Lehman Brothers in 2008 have led to periods with changes in mean, volatility and correlations in stock market returns [15], stressing the non-stationary nature and the existence of drastic structural changes in financial markets [18], [20]–[23].

In the financial literature, changes in the price behaviour of financial markets that go beyond their normal price fluctuations receive the name of regime changes [19], [53],

[63] or business cycles shifts [80]. In order to model these regime changes, one of the most popular techniques is the regime-switching model [15], which was first applied by Hamilton [58] as a technique to deal with cycles of different economic activities such as recessions and market expansions.

In financial markets, there are periods of time with different degrees of efficiency and predictability. There can be moments where, due to the market-wide sentiment given by political or economic circumstances, the behaviour of investors may change towards a bear, bull, lateral market, and periods or time frames with different levels of volatility [80].

At the macroeconomic level, RC are often related to abrupt breaks in long-term cycles like the break of bubbles or economic crises [59]. Changes in market regimes could be driven as well by investor expectations [15]. The financial literature identifies two types of regimes clear to recognise: steady and highly volatile regimes usually linked to economic growth or deflation periods, respectively.

This is illustrated in Fig. 8, which shows the breaks identified in [19] during the Great Recession.

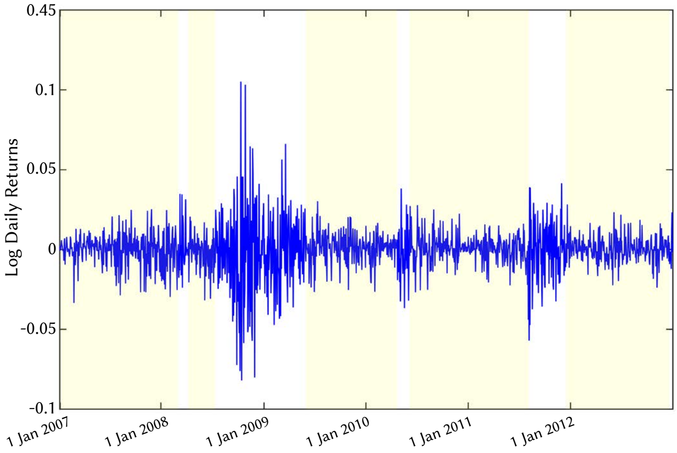


Fig. 8. Regime Changes in the DJIA Index (indicator of the United States economy) identified by Tsang and Chen [19].

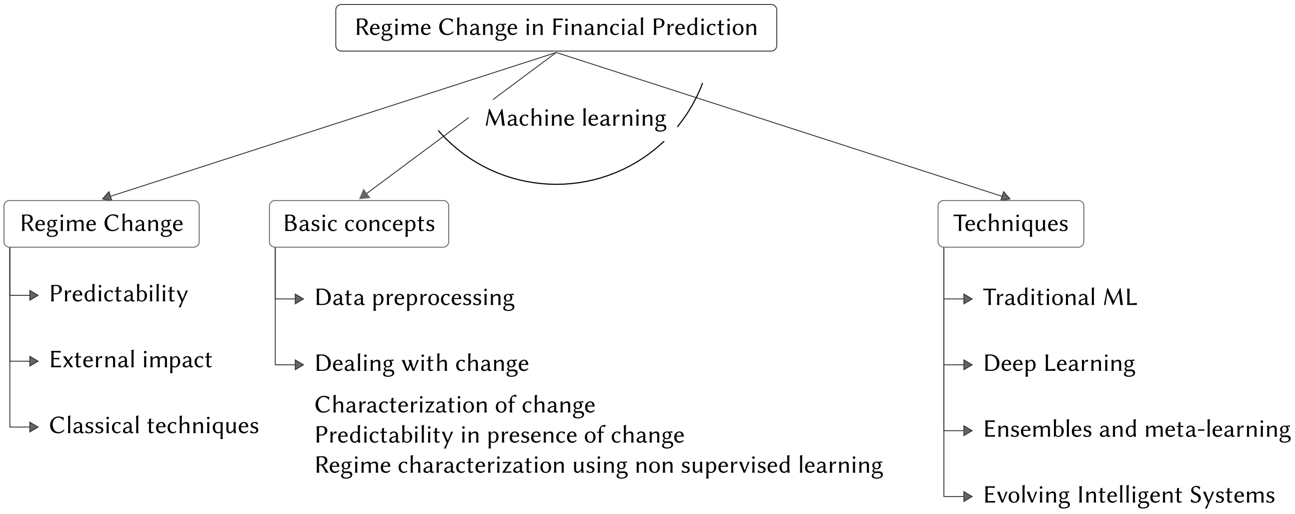

Regime Change in Financial Prediction
Machine learning
Regime Change
Predictability
External impact
Classical techniques
Basic concepts
Data preprocessing
Dealing with change
Characterization of change
Predictability in presence of change
Regime characterization using non supervised learning
Techniques
Traditional ML
Deep Learning
Ensembles and meta-learning
Evolving Intelligent Systems

Fig. 7. Discussion on the research papers considered in this work.

Regime changes challenge investors, making them change their trading strategies as the collective trading behaviour of the market changes. Different examples of RC have been covered in the recent literature. Davies [53] analysed different cases and consequences of regime changes in the Great Recession that impacted several asset classes such as equities, bonds, commodities and currencies at micro and macroeconomic levels. Hamilton [63] observed alternating patterns between steady and turbulent periods since the Second World War and subsequent recessions by looking at US unemployment rates. Ang and Timmermann [15] identified cyclic changes in the behaviour of asset prices and mean, volatility and correlation patterns in stock returns during the Great Recession and the 1973 oil crisis. Kritzman et al. [29] discovered that investors could benefit from having different asset allocation strategies in different market regimes to minimise losses.

Many other studies consider these drastic changes an intrinsic characteristic of financial data that might be caused by significant events, and thus, these will be observable not only in prices and economic variables but also in other kinds of public information [52], [58], [59], [62], [63], [92]. Hamilton [58] proposed a time-series based approach [26] to capture nonlinear effects like RC, identify market breaks and hidden changes in economic cycles known as the regime-switching model [59]. This model, also known as the Markov-switching model, is fitted to observations following different patterns in different periods and is mainly applied to recognise low volatility regimes with economic growth vs high volatility periods with economic contractions [116]. Ang and Timmermann [15] applied these models to predict interest rates and equity and foreign exchange returns. They discussed how to model RCs for these time series models.

B. Approaches Based on Machine Learning (Q2)

Traditional statistical methods tend to model and predict future data based on the assumption that the time series under study is generated from a linear process with features normally distributed. This presents challenges since financial data is characterised by nonlinearity and non-stationarity besides a high level of uncertainty and noise [82].

A different approach to performing financial forecasts is the use of ML techniques. Several literature reviews show the benefits of these techniques against traditional methods [5], [98], [99], [105], surprising practitioners by contradicting early theories like the random walk, and efficient market hypothesis (EMH) [5], [25], [95]. Machine learning algorithms can handle nonlinear relationships without prior knowledge [7], [144], outperforming traditional time series methods [10]–[13].

Different research works from the economic literature have either used technical indicators or raw prices and returns. Technical indicators are able to show behavioural patterns among traders and thus provide an extra level of signal to predictive models. These can be valuable to automate the behaviour of short-term traders [5]. In any case, most of the economics literature has focused on linear processes that may not have been able to extract relevant information nor infer complex relationships among technical indicators where some new ML methods could [108].

Some of the literature reviews already cited describe common technical indicators used for stock market value and trend prediction. Many of these papers, like [7], have shown that different pre-processing steps, like the frequency level of the input data, can impact its predictability. While a common approach in this regard is data normalisation, in the literature on data stream mining, data normalisation is not a usual practice since maximum and minimum values for each attribute in the data stream are unknown beforehand [134].

Authors like Patel et al. [13] discretise features based on the human approach to investing and deriving the technical indicators using assumptions from the stock market. This latter approach, though, introduces human bias in the process. This is the opposite way to approach the problem of trend prediction if we compare it to recent deep learning strategies that feed dozens of automatically generated indicators [109].

Overall, the above-mentioned literature reviews confirm that ML techniques can be used to predict price changes, but this entirely depends on the time horizon and efficiency of the market in the period predicted. Cavalcante et al. [3] provided another interesting review of pre-processing and clustering techniques used in the financial domain to forecast future market movements. They highlighted the relevance of concept drifts in financial markets and suggested that the data stream mining literature is of great importance in future research due to the non-stationarity and evolution of financial markets [91].

In computational finance, changes in the behaviour of the market are normally referred to as regime changes or switches [15], [61], [67], structural breaks or changes [51], [54], [56], [56], volatility shifts [50], switching processes [60] or market states [18]. In this kind of data, long periods of stability might be interrupted by short episodes of abrupt changes [61].

These changes may or not be transitory since a newly adopted behaviour in price dynamics, reflected as part of the mean returns, their volatility or correlation among them may persist for several periods. Timely recognition of these sudden behavioural changes in markets can significantly lower the risk of financial exposure. This has inspired the materialisation of techniques such as regime-switching models in the financial literature, which work under the premise that new dynamics of price returns and fundamentals persist for several periods after a change. A key element in these models is identifying whether the exact market regimes reoccur over time (e.g. across recessions or periods of economic growth) or if new regimes deviate or have evolved from previous ones [15].

The prediction of future values in financial markets and the detection of regime changes in data streams with temporal dependence are common application areas for statistical methods and ML. Previous research reports high accuracy in forecasting price changes with advanced techniques and the feasibility of making profits using these predictions against the EMH, which points to unbeatable markets. An alternative theory is the adaptive market hypothesis (AMH) [57], introduced by Andrew Lo in 2004. This theory, with empirical evidence in a increasing number of research works [68], combines the EMH with principles from behavioural finance, allowing the ideas of market efficiency and inefficiencies to co-exist. Under the AMH, the efficiency of a market evolves as market participants adapt to an environment that changes continuously. In this regard, participants rely on heuristics to make their investment choice, leading to mostly rational markets under those heuristics (like the EMH). The main difference is at the time of major behavioural shifts in the market participants, as in economic shocks or crises. In this case, the AMH considers a market that evolves, and the initially adaptive heuristics may become static in certain market situations. Consequently, the EMH may not continue under periods of abnormal conditions, stress or abrupt changes in the market. Hence, financial markets may be predictable in specific periods, as discussed by Lo [49]. Therefore, convergence to market efficiency is neither guaranteed nor likely to occur. The level of efficiency depends on the market participants and the market conditions at that time.

One of the few financial studies citing concept drift explicitly can be found in a recent work authored by Masegosa et al. [36]. They analysed data from the Great Recession and claimed that economic changes during this period manifested as concept drifts in their generative processes. An intermediate example of trying to predict financial crises using ML methods can be found in [55], where the authors studied possible contagion risks between financial markets that could trigger financial crises to signal warnings at an early stage. More recently, Yang et al. [137] analysed the impact of concept drift in business processes. More specifically, they modelled the response to concept drift as a sequential decision-making problem by combing a hierarchical Markov model and a Markov decision process. Martín et al. [145] also dealt with structural changes introducing a trading system based on grammatical evolution that commutes between an active model and a candidate one to increase performance.

Other two key research pieces in this regard are the works by Tsang and Chen [19] and M "unnix et al. [18], which proposed mechanisms to identify points of drastic changes in financial time series. The former used statistical-based and traditional ML (e.g. naive Bayes) approaches to classify normal versus abnormal regimes. They proposed a framework based on the change speed of price returns and the degree of changes to visualise and discriminate between different market regimes depending on the volatility of their price returns. The latter [18] visualised differences in the correlation structure of the price returns across assets in the S&P500 during the Great Recession. They extended the selection to a sample from 1992 to 2010, identifying eight market states repeating behavioural changes over time.

Several approaches from the deep learning (DL) field (e.g. RNNs) have also tried to face the problem of concept changes when learning continuously. [113], [126], [128], [130]. These advances have been successfully transferred to the financial domain, as discussed in recent surveys by Ozbayoglu et al. [9], and Li and Bastos [109].

Most of these mentioned research works focus on the likelihood of daily or monthly changes, where retraining a model is a feasible task. As of today, the amount of research devoted to seasonality and changes at the intraday level is significantly more limited [75], [89], [101], [103], [107]. The computational cost of ML and statistical methods, together with the inherent higher complexity derived from the need to manage large amounts of data at these resolutions, makes keeping models up to date more challenging.

While its application to high-frequency markets is still an open problem, recent research works are making progress in understanding how to apply ML to intraday resolutions. Among them, we could mention the one presented by Sirignano and Cont [73], who claimed that financial data at high frequencies exhibit stylised facts and may hold learnable stationary patterns over long periods. Another relevant study is the one authored by Shintate and Pichl [76], who reviewed modern ML and DL approaches applied to high-frequency trading at the minute level.

Recently, several research works have approached the problem of time-changing behaviours using non-supervised ML methods [127], [130], [135]. In these, micro-clusters or latent features may be used to represent a summary of the incoming data and reduce the computational costs of correlating full data distributions. A manner of doing this is using model-based clustering approaches. These algorithms find models that fit input data and are also robust to the presence of noise [146], [147]. For instance, expectation maximisation (EM) [148] fits a mixture of Gaussian distributions to the data [133]. Chiu and Minku [141] used it in concept drift handling based on clustering in the model space (CDCMS) to create concept representations and keep a diverse ensemble learner. Zheng et al. [123] relied on it to minimise intra-cluster dispersion and cluster impurity. Tsang and Chen [19] applied the Baum–Welch algorithm, a special case of EM, to both detect the time of a change point and predict the next state (or concept) of financial data using a hidden Markov model (HMM). Gomes et al. [122] also hypothesised about using Baum–Welch in conjunction with HMMs for continuous learning problems. Baum–Welch has been used in the financial domain together with other specific versions of EM and Gaussian mixture models (GMM) to forecast change direction in stock prices [72], [87] and to represent market regimes [19], [29], [64], [86].

A set of relevant techniques from the ML literature in this regard are prototype generation techniques such as learning vector quantisation [65], [96], [102], which have been proven to be useful for data partitioning and model selection in the financial domain. Choudhury et al. [71], and Pavlidis et al. [69] use a combination of clustering and forecasting algorithms to model the distribution of financial data. Regarding the first step, the former authors use a two-layer abstraction that clusters stocks using self-organising maps (SOM) to then rely on K-means to obtain clusters of prototypes. The latter considers three different unsupervised algorithms to identify market states: growing neural gas (GNG), density-based spatial clustering of applications with noise (DBSCAN), and Unsupervised k-Windows. Once the market states are identified, they use feed-forward neural networks to make predictions.

In the last years, several online incremental algorithms have used these techniques to adapt distinct learners to different cycles or seasonal behaviours in a data stream. In this regard, the use of online ensembles, using non-supervised learning to represent different behavioural patterns [135], [141] or supervised learning to train a pool of classifiers with high predictive accuracy under different conditions [34], [117], have obtained state-of-the-art results adapting to different states in the behaviour of data streams in many domains, including finance [17].

Meta learners over data streams are a related subfield of ML of increasing popularity where a pool of former classifiers is managed and reused when the state or concept of the stream changes. This subfield is inspired by the human learning system that reuses previous knowledge to learn new tasks, not starting from zero every time. Although meta-learners have not been widely applied to the financial domain yet, their logic resembles approaches applied to finance as EM or Baum-Welch, but for continuous learning domains where models are always up-to-date and thus behave smoothly in case of structural breaks. For instance, Abad et al. [120] proposed a meta-learner that used hidden Markov models (HMM) to predict the sequence of change between discrete concepts. Their approach used fuzzy logic rules to compare classifiers to reuse former models. Maslov et al. [139] proposed a method to use patterns acquired during previous changes and assumed a Gaussian distribution for the duration of the changes to predict the time of the next change point. Carta et al. in [83] recently combined the use of meta learners with deep reinforcement learning to produce trading strategies and maximise profits operating in Standard & Poor's 500 future markets and the J.P. Morgan and Microsoft stocks.

Meta-learning approaches have inbuilt strategies to decide on when to train, what model to replace and when, when to forget (prune) a learner and when to create one [114], [136] by using the evaluation performance metrics of active and former models [149]. These, thus, are a closely related research area to evolving intelligent systems (EIS) [118], [129], [150] and evolving fuzzy

systems [125] [132] [143] [119]. EIS, which are also online and incremental systems, can adapt themselves to concept drifts of different natures on-the-fly through adaptive fuzzy rules [140]. EIS have already demonstrated their ability to solve different kinds of problems in various application domains like finance [90], [97], [151]. These have achieved great results in classifying non-stationary time series [24], [37], [38].

Recent EIS approaches can work as ensembles of rules [116] and apply meta-cognitive scaffolding theory for tuning the learned model incrementally in what-to-learn, when-to-learn, and how-to-learn [138]. In fact, ensembles are known for their good results in predicting both cyclic and non-stationary data such as stock prices [10], [13], [110]. These have also introduced the ability to deal with recurrent concepts explicitly and have beaten other methods at predicting the S&P500 [37], [38], [111]. For instance, Pratama et al. employed an evolving type-2 recurrent fuzzy neural network to learn incrementally and handle recurring drifts in both [124] and [38]. In any case, there is still a significant gap between EIS and the rest of the literature for data stream classification.

This line of work based on EIS is being complemented by other studies that combine ensembles and other evolutionary algorithms to tackle concept drift in financial applications like [84]. These authors used ensembles of trading rules evolved using grammatical evolution to manage structural change in the Standard & Poor's 500 index.

V. Conclusions and Future Work

The application of AI to computational finance has been a very active field of research for decades. Among the key difficulties identified in the literature on financial prediction, we can mention structural change. The price generation process of financial time series is often affected by changes of different natures. Some of these changes can reoccur over time as seasonal patterns, while others do not repeat, being abrupt breaks in the non-stationary price dynamics.

This study presented a systematic literature review of machine learning techniques for financial prediction under regime changes. A variety of sources were inspected to perform an exhaustive search. This review included: ScienceDirect, IEEE Digital Library, ACM Digital Library, Taylor & Francis, Wiley and SpringerLink. Out of a total of 140 relevant studies, these are distributed as follows: i) concept drift or online learning related (32.1%); ii) related to financial literacy for regime changes (15.7%), and iii) ML techniques applied to stock forecasting (52.1%).

The results of reviewed publications show that ML has proven to be a powerful tool to tackle the problem of financial prediction under concept drift, which we define as structural breaks, that can occur at any frequency level. This includes solutions based on different algorithms that adapt their prediction models to new circumstances either through new model generation or managing an archive of former successful models. In this regard, many meta-learning approaches in the ML literature rely on non-supervised algorithms to try to identify the recurrence of a concept and retrieve previous models or detect drifts.

In this context, the use of sequential DL models such as RNNs can be insufficient to tackle abrupt changes since the previous market dynamics are still in memory in the models impacting predictive accuracy. In contrast, the model learns the new regime [113], [115], [121].

Depending on the frequency involved, researchers suggest either solutions based on model retraining either at regular intervals or upon detection of changes or drifts or online incremental algorithms. This entails having up-to-date models with the use of forgetting mechanisms to avoid overfitting and adapting to new market behaviours. Regarding the latter, and despite the success of online ensembles dealing with complex systems and training base learners to deal with different regimes, the use of these approaches from the data stream learning literature is not as popular in financial forecasting yet.

Even though there doesn't seem to be a clearly dominant ML technique in this space, it is worth mentioning the popularity of solutions based on ensembles and evolving fuzzy systems. It is also important to note how the relatively recent developments in deep learning have fostered the popularity of approaches where artificial neural networks play a key role.

Future research is likely to emphasise the application of data stream classification algorithms to financial streams. Online machine learning has not been widely applied to the financial domain. However, as shown in this study, similar techniques like sequential and recurring deep learning models are on the rise in finance. Applying the problem of concept drift to handling price change dynamics seems a natural step forward on the research line of financial regime changes.

Having said that, better access to high-frequency data and computational resources will also likely result in major progress in the near future.

Acknowledgment

We would like to thank the editor and external reviewers for their thoughtful and detailed comments on our paper. We would also like to acknowledge the financial support of the Spanish Ministry of Science, Innovation and Universities under grant PGC2018-096849-B-I00 (MCFin).

References

[1] Y. S. Abu-Mostafa, A. F. Atiya, "Introduction to financial forecasting," *Applied Intelligence*, vol. 6, pp. 205–213, 7 1996.

[2] W. Huang, Y. Nakamori, S.-Y. Wang, "Forecasting stock market movement direction with support vector machine," *Computers & Operations Research*, vol. 32, pp. 2513–2522, 2005.

[3] R. C. Cavalcante, R. C. Brasileiro, V. L. F. Souza, J. P. Nobrega, A. L. I. Oliveira, "Computational Intelligence and Financial Markets: A Survey and Future Directions," *Expert Systems with Applications*, vol. 55, pp. 194–211, 8 2016.

[4] V. Dogra, S. Verma, Kavita, N. Z. Jhanjhi, U. Ghosh, D. N. Le, "A Comparative Analysis of Machine Learning Models for Banking News Extraction by Multiclass Classification With Imbalanced Datasets of Financial News: Challenges and Solutions," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 3, pp. 35–52, 2022, doi: 10.9781/ijimai.2022.02.002.

[5] M. W. Hsu, S. Lessmann, M. C. Sung, T. Ma, J. E. Johnson, "Bridging the divide in financial market forecasting: machine learners vs. financial economists," *Expert Systems with Applications*, vol. 61, pp. 215–234, 2016.

[6] B. M. Henrique, V. A. Sobreiro, H. Kimura, "Literature review: Machine learning techniques applied to financial market prediction," *Expert Systems with Applications*, vol. 124, pp. 226–251, 2019.

[7] G. S. Atsalakis, K. P. Valavanis, "Surveying stock market forecasting techniques - Part II: Soft computing methods," *Expert Systems with Applications*, vol. 36, pp. 5932-59-41, 4 2009.

[8] M.-Y. Chen, A. K. Sangaiah, T.-H. Chen, E. D. Lughofer, E. Egrioglu, "Deep learning for financial engineering," *Computational Economics*, pp. 1–5, 2022.

[9] A. M. Ozbayoglu, M. U. Gudelek, O. B. Sezer, "Deep learning for financial applications: A survey," *Applied Soft Computing*, vol. 93, p. 106384, 2020, doi: https://doi.org/10.1016/j.asoc.2020.106384.

[10] M. Ballings, D. Van Den Poel, N. Hespeels, R. Gryp, "Evaluating multiple classifiers for stock price direction prediction," *Expert Systems with Applications*, vol. 42, no. 20, pp. 7046–7056, 2015.

[11] A. Booth, E. Gerding, F. McGroarty, "Automated trading with performance weighted random forests and seasonality," *Expert Systems with Applications*, vol. 41, pp. 3651–3661, 6 2014.

[12] P. Ładyżyński, K. Żbikowski, P. Grzegorzewski, "Stock trading with random forests, trend detection tests and force index volume indicators," in *Artificial Intelligence and Soft Computing: 12th International Conference, ICAISC 2013, Proceedings, Part II*, Berlin, Heidelberg, Jun. 2013, pp. 441–452, Springer Berlin Heidelberg.

[13] J. Patel, S. Shah, P. Thakkar, K. Kotecha, "Predicting stock and stock price

index movement using trend deterministic data preparation and machine learning techniques," *Expert Systems with Applications*, vol. 42, no. 1, pp. 259–268, 2015, doi: https://doi.org/10.1016/j.eswa.2014.07.040.

[14] D. Ardia, K. Bluteau, M. Rüede, "Regime changes in Bitcoin GARCH volatility dynamics," *Finance Research Letters*, vol. 29, pp. 266–271, Jun. 2019, doi: 10.1016/J.FRL.2018.08.009.

[15] A. Ang, A. Timmermann, "Regime changes and financial markets," *Annual Review of Financial Economics*, vol. 4, no. 1, pp. 313–337, 2012.

[16] A. Tsymbal, "The Problem of Concept Drift: Definitions and Related Work," *Technical Report: TCD-CS-2004-15, Department of Computer Science Trinity College, Dublin*, 2004.

[17] A. L. Suárez-Cetrulo, A. Cervantes, D. Quintana, "Incremental Market Behavior Classification in Presence of Recurring Concepts," *Entropy*, vol. 21, p. 25, Jan. 2019, doi: 10.3390/e21010025.

[18] M. C. Münnix, T. Shimada, R. Schäfer, F. Leyvraz, T. H. Seligman, T. Guhr, H. E. Stanley, "Identifying States of a Financial Market," *Scientific Reports*, vol. 2, p. 644, 12 2012.

[19] E. Tsang, J. Chen, *Detecting regime change in computational finance: data science, machine learning and algorithmic trading*. CRC Press, 2020.

[20] R. T. Das, K. K. Ang, C. Quek, "IeRSPOP: A novel incremental rough set-based pseudo outer-product with ensemble learning," *Applied Soft Computing Journal*, vol. 46, pp. 170–186, 9 2016.

[21] V. Vella, W. L. Ng, "Enhancing risk-adjusted performance of stock market intraday trading with Neuro-Fuzzy systems," *Neurocomputing*, vol. 141, pp. 170–187, 2014.

[22] Y. Hu, K. Liu, X. Zhang, K. Xie, W. Chen, Y. Zeng, M. Liu, "Concept drift mining of portfolio selection factors in stock market," *Electronic Commerce Research and Applications*, vol. 14, no. 6, pp. 444–455, 2015.

[23] B. Silva, N. Marques, G. Panosso, "Applying neural networks for concept drift detection in financial markets," in *CEUR Workshop Proceedings*, vol. 960, 2012, pp. 43–47.

[24] X. Gu, P. P. Angelov, A. M. Ali, W. A. Gruver, G. Gaydadjiev, "Online evolving fuzzy rule-based prediction model for high frequency trading financial data stream," in *2016 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS)*, 5 2016, pp. 169–175, IEEE.

[25] E. F. Fama, "Efficient Capital Markets: A Review of Theory and Empirical Work," *The Journal of Finance*, vol. 25, no. 2, p. 383, 1970, doi: 10.2307/2325486.

[26] J. Piger, *Econometrics: Models of Regime Changes*, pp. 2744–2757. New York, NY: Springer New York, 2009.

[27] A. G. Hoepner, D. McMillan, A. Vivian, Wese Simen, "Significance, relevance and explainability in the machine learning age: an econometrics and financial data science perspective," *The European Journal of Finance*, vol. 27, no. 1-2, pp. 1–7, 2021.

[28] P. Bracke, A. Datta, C. Jung, S. Sen, "Machine learning explainability in finance: an application to default risk analysis," Bank of England, 2019.

[29] M. Kritzman, S. Page, D. Turkington, "Regime shifts: Implications for dynamic strategies (corrected)," *Financial Analysts Journal*, vol. 68, no. 3, pp. 22–39, 2012.

[30] R. Elwell, R. Polikar, "Incremental learning of concept drift in nonstationary environments.," *IEEE transactions on neural networks*, vol. 22, pp. 1517–31, 10 2011.

[31] J. A. Gama, I. Žliobaitundefined, A. Bifet, M. Pechenizkiy, A. Bouchachia, "A survey on concept drift adaptation," *ACM Computing Surveys*, vol. 46, Mar. 2014, doi: 10.1145/2523813.

[32] G. Ditzler, M. Roveri, C. Alippi, R. Polikar, "Learning in nonstationary environments: A survey," *IEEE Computational Intelligence Magazine*, vol. 10, pp. 12–25, Nov. 2015.

[33] G. I. Webb, R. Hyde, H. Cao, H. L. Nguyen, F. Petitjean, "Characterizing concept drift," *Data Mining and Knowledge Discovery*, vol. 30, pp. 964–994, 7 2016.

[34] H. M. Gomes, J. P. Barddal, F. Enembreck, A. Bifet, "A Survey on Ensemble Learning for Data Stream Classification," *ACMComputing Surveys*, vol. 50, no. 2, pp. 1–36, 2017, doi: 10.1145/3054925.

[35] S. Ramírez-Gallego, B. Krawczyk, S. García, M. Woźniak, F. Herrera, "A survey on data preprocessing for data stream mining: Current status and future directions," *Neurocomputing*, vol. 239, pp. 39–57, 2017.

[36] A. R. Masegosa, A. M. Martínez, D. Ramos-López, H. Langseth, T. D. Nielsen, A. Salmerón, "Analyzing concept drift: A case study in the financial sector," *Intelligent Data Analysis*, vol. 24, no. 3, pp. 665–688, 2020.

[37] M. Pratama, E. Lughofer, J. Er, S. Anavatti, C.-P. Lim, "Data driven modelling based on Recurrent Interval-Valued Metacognitive Scaffolding Fuzzy Neural Network," *Neurocomputing*, vol. 262, pp. 4–27, 2017.

[38] M. Pratama, J. Lu, E. Lughofer, G. Zhang, M. J. Er, "Incremental Learning of Concept Drift Using Evolving Type-2 Recurrent Fuzzy Neural Network," *IEEE Transactions on Fuzzy Systems*, pp. 1–1, 2016, doi: 10.1109/TFUZZ.2016.2599855.

[39] C. Alippi, G. Boracchi, M. Roveri, "Just-in-time classifiers for recurrent concepts," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, pp. 620–634, 4 2013.

[40] J. B. Gomes, M. M. Gaber, P. A. C. Sousa, E. Menasalvas, "Mining recurring concepts in a dynamic feature space," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, pp. 95–110, 1 2014, doi: 10.1109/TNNLS.2013.2271915.

[41] P. M. Gonçalves Jr, R. Souto, M. De Barros, "RCD: A recurring concept drift framework," *Pattern Recognition Letters*, vol. 34, pp. 1018–1025, 2013.

[42] A. W. Lo, H. Mamaysky, J. Wang, "Foundations of Technical Analysis: Computational Algorithms, Statistical Inference, and Empirical Implementation," *The Journal of Finance*, vol. 55, pp. 1705–1765, 8 2000.

[43] F. E. Tay, L. Cao, "Application of support vector machines in financial time series forecasting," *Omega*, vol. 29, pp. 309–317, 8 2001.

[44] T. Geva, J. Zahavi, "Empirical evaluation of an automated intraday stock recommendation system incorporating both market data and textual news," *Decision Support Systems*, vol. 57, pp. 212–223, Jan. 2014, doi: 10.1016/J.DSS.2013.09.013.

[45] C. H. Chen, P. Y. Chen, J. C. W. Lin, "An Ensemble Classifier for Stock Trend Prediction Using Sentence- Level Chinese News Sentiment and Technical Indicators," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 3, pp. 53–64, 2022, doi: 10.9781/ijimai.2022.02.004.

[46] B. Kitchenham, O. P. Brereton, D. Budgen, M. Turner, J. Bailey, S. Linkman, "Systematic literature reviews in software engineering–a systematic literature review," *Information and software technology*, vol. 51, no. 1, pp. 7– 15, 2009.

[47] B. Kitchenham, "Procedures for performing systematic reviews," *Keele, UK, Keele University*, vol. 33, no. 2004, pp. 1–26, 2004.

[48] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, *et al.*, "The prisma 2020 statement: an updated guideline for reporting systematic reviews," *Systematic reviews*, vol. 10, no. 1, pp. 1–11, 2021.

[49] A. W. Lo, "Reconciling efficient markets with behavioral finance: the adaptive markets hypothesis," *Journal of investment consulting*, vol. 7, no. 2, pp. 21–44, 2005.

[50] S. Baek, S. K. Mohanty, M. Glambosky, "Covid-19 and stock market volatility: An industry level analysis," *Finance Research Letters*, vol. 37, p. 101748, 2020.

[51] M. L. De Prado, *Advances in financial machine learning*. John Wiley & Sons, 2018.

[52] A. Ang, G. Bekaert, "How regimes affect asset allocation," *Financial Analysts Journal*, vol. 60, no. 2, pp. 86–99, 2004.

[53] G. Davies, "Regime changes in the financial markets," 2016. [Online]. Available: https://www.ft.com/content/6556ec60-6aa3-3dfe- 8953-b94d6080c360.

[54] E. Andreou, E. Ghysels, *Structural Breaks in Financial Time Series*, pp. 839–870. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009.

[55] A. Samitas, E. Kampouris, D. Kenourgios, "Machine learning as an early warning system to predict financial crisis," *International Review of Financial Analysis*, vol. 71, p. 101507, 2020.

[56] D. Pettenuzzo, A. Timmermann, "Predictability of stock returns and asset allocation under structural breaks," *Journal of Econometrics*, vol. 164, pp. 60–78, Sep. 2011, doi: 10.1016/j.jeconom.2011.02.019.

[57] A. W. Lo, "The Adaptive Markets Hypothesis: Market Efficiency from an Evolutionary Perspective," MIT, 2004.

[58] J. D. Hamilton, "A new approach to the economic analysis of nonstationary time series and the business cycle," *Econometrica: Journal of the econometric society*, pp. 357–384, 1989.

[59] J. D. Hamilton, "Regime switching models," in *Macroeconometrics and time series analysis*, Springer, 2010, pp. 202–209.

[60] T. Preis, J. J. Schneider, H. E. Stanley, "Switching processes in financial markets," *Proceedings of the National Academy of Sciences*, vol. 108, no. 19, pp. 7674–7678, 2011, doi: 10.1073/pnas.1019484108/-/DCSupplemental.

[61] R. Hammerschmid, H. Lohre, “Regime shifts and stock return predictability,” *International Review of Economics and Finance*, vol. 56, pp. 138–160, Jul. 2018, doi: 10.1016/j.iref.2017.10.021.

[62] Q. Dai, K. J. Singleton, W. Yang, “Is regime-shift risk priced in the us treasury market?,” Working paper, New York University and Stanford University, 2003.

[63] J. D. Hamilton, “Macroeconomic regimes and regime shifts,” *Handbook of macroeconomics*, vol. 2, pp. 163– 201, 2016.

[64] J. G. Dias, J. K. Vermunt, S. Ramos, “Clustering financial time series: New insights from an extended hidden markov model,” *European Journal of Operational Research*, vol. 243, no. 3, pp. 852–864, 2015.

[65] K.-S. Kim, I. Han, “The cluster-indexing method for case-based reasoning using self-organizing maps and learning vector quantization for bond rating cases,” *Expert systems with applications*, vol. 21, no. 3, pp. 147– 156, 2001.

[66] R. Bisoi, P. K. Dash, “A hybrid evolutionary dynamic neural network for stock market trend analysis and prediction using unscented kalman filter,” *Applied Soft Computing*, vol. 19, pp. 41–56, 2014.

[67] M. Guidolin, A. Timmermann, “An econometric model of nonlinear dynamics in the joint distribution of stock and bond returns,” *Journal of Applied Econometrics*, vol. 21, pp. 1–22, Jan. 2006, doi: 10.1002/jae.824.

[68] A. Urquhart, F. McGroarty, “Are stock markets really efficient? evidence of the adaptive market hypothesis,” *International Review of Financial Analysis*, vol. 47, pp. 39–49, 2016, doi: https://doi.org/10.1016/j.irfa.2016.06.011.

[69] N. G. Pavlidis, V. P. Plagianakos, D. K. Tasoulis, M. N. Vrahatis, “Financial forecasting through unsupervised clustering and neural networks,” *Operational Research*, vol. 6, no. 2, pp. 103–127, 2006.

[70] A. A. Ariyo, A. O. Adewumi, C. K. Ayo, “Stock price prediction using the ARIMA model,” in *2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation*, 2014, pp. 106–112, IEEE.

[71] S. Choudhury, S. Ghosh, A. Bhattacharya, K. J. Fernandes, M. K. Tiwari, “A real time clustering and SVM based price-volatility prediction for optimal trading strategy,” *Neurocomputing*, vol. 131, pp. 419– 426, 5 2014.

[72] S.-H. Park, J.-H. Lee, J.-W. Song, T.-S. Park, “Forecasting change directions for financial time series using hidden Markov model,” in *International Conference on Rough Sets and Knowledge Technology*, 2009, pp. 184–191, Springer.

[73] J. Sirignano, R. Cont, “Universal features of price formation in financial markets: perspectives from deep learning,” *Quantitative Finance*, 2019, doi: 10.1080/14697688.2019.1622295.

[74] G. Deboeck, *Trading on the edge: neural, genetic, and fuzzy systems for chaotic financial markets.* Wiley, 1994.

[75] S. Schulmeister, “Profitability of technical stock trading: Has it moved from daily to intraday data?,” *Review of Financial Economics*, vol. 18, no. 4, pp. 190– 201, 2009, doi: 10.1016/j.rfe.2008.10.001.

[76] T. Shintate, L. Pichl, “Trend Prediction Classification for High Frequency Bitcoin Time Series with Deep Learning,” *Journal of Risk and Financial Management*, vol. 12, p. 17, Jan. 2019, doi: 10.3390/jrfm12010017.

[77] B. G. Malkiel, “The efficient market hypothesis and its critics,” *Journal of economic perspectives*, vol. 17, no. 1, pp. 59–82, 2003.

[78] A. W. Lo, A. C. MacKinlay, *A Non-Random Walk Down Wall Street.* Princeton University Press, Dec. 2011.

[79] X. Ding, Y. Zhang, T. Liu, J. Duan, “Using structured events to predict stock price movement: An empirical investigation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1415–1425.

[80] F. Black, “Noise,” *The Journal of Finance*, vol. 41, 529–543, Jul. 1986, doi: 10.1111/j.1540- 6261.1986.tb04513.x.

[81] B. Malkiel, *A Random Walk Down Wall Street.* WW Norton & Company, 1973.

[82] B. Vanstone, G. Finnie, “An empirical methodology for developing stockmarket trading systems using artificial neural networks,” *Expert Systems with Applications*, vol. 36, no. 3, Part 2, pp. 6668–6680, 2009, doi: https://doi.org/10.1016/j.eswa.2008.08.019.

[83] S. Carta, A. Corriga, A. Ferreira, A. S. Podda, D. R. Recupero, “A multi-layer and multi-ensemble stock trader using deep learning and deep reinforcement learning,” *Applied Intelligence*, vol. 51, no. 2, pp. 889–905, 2021.

[84] C. Martín, D. Quintana, P. Isasi, “Grammatical evolution-based ensembles for algorithmic trading,” *Applied Soft Computing*, vol. 84, p. 105713, 2019, doi: https://doi.org/10.1016/j.asoc.2019.105713.

[85] C. Chen, W. Dongxing, H. Chunyan, Y. Xiaojie, “Exploiting social media for stock market prediction with factorization machine,” in *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, vol. 2, 2014, pp. 142–149, IEEE.

[86] M. R. Hassan, K. Ramamohanarao, J. Kamruzzaman, M. Rahman, M. M. Hossain, “A hmm-based adaptive fuzzy inference system for stock market forecasting,” *Neurocomputing*, vol. 104, pp. 10–25, 2013.

[87] X. Zhang, Y. Li, S. Wang, B. Fang, S. Y. Philip, “Enhancing stock market prediction with extended coupled hidden Markov model over multi-sourced data,” *Knowledge and Information Systems*, vol. 61, no. 2, pp. 1071–1090, 2019.

[88] N. Gârleanu, L. H. Pedersen, “Efficiently inefficient markets for assets and asset management,” *The Journal of Finance*, vol. 73, no. 4, pp. 1663–1712, 2018.

[89] G. C. Friesen, P. A. Weller, L. M. Dunham, “Price trends and patterns in technical analysis: A theoretical and empirical examination,” *Journal of Banking and Finance*, vol. 33, pp. 1089–1100, Jun. 2009, doi: 10.1016/j.jbankfin.2008.12.010.

[90] C. F. Liu, C. Y. Yeh, S. J. Lee, “Application of type- 2 neuro-fuzzy modeling in stock price prediction,” *Applied Soft Computing Journal*, vol. 12, no. 4, pp. 1348– 1358, 2012, doi: 10.1016/j.asoc.2011.11.028.

[91] R. C. Cavalcante, A. L. Oliveira, “An autonomous trader agent for the stock market based on online sequential extreme learning machine ensemble,” in *2014 International Joint Conference on Neural Networks (IJCNN)*, 2014, pp. 1424–1431, IEEE.

[92] R. Garcia, R. Luger, E. Renault, “Empirical assessment of an intertemporal option pricing model with latent variables,” *Journal of Econometrics*, vol. 116, no. 1-2, pp. 49–83, 2003.

[93] R. Dash, P. K. Dash, R. Bisoi, “A self adaptive differential harmony search based optimized extreme learning machine for financial time series prediction,” *Swarm and Evolutionary Computation*, vol. 19, pp. 25– 42, 2014.

[94] A. A. Adebiyi, C. K. Ayo, M. O. Adebiyi, S. O. Otokiti, “Stock price prediction using neural network with hybridized market indicators,” *Journal of Emerging Trends in Computing and Information Sciences*, vol. 3, no. 1, pp. 1–9, 2012.

[95] R. Tsaih, Y. Hsu, C. C. Lai, “Forecasting S&P 500 stock index futures with a hybrid AI system,” *Decision Support Systems*, vol. 23, pp. 161–174, 6 1998.

[96] J. C. Neves, A. Vieira, “Improving bankruptcy prediction with hidden layer learning vector quantization,” *European Accounting Review*, vol. 15, no. 2, pp. 253–271, 2006.

[97] Y. L. Yong, Y. Lee, X. Gu, P. P. Angelov, D. C. L. Ngo, E. Shafipour, “Foreign currency exchange rate prediction using neuro-fuzzy systems,” *Procedia computer science*, vol. 144, pp. 232–238, 2018.

[98] H. Ghoddusi, G. G. Creamer, N. Rafizadeh, “Machine learning in energy economics and finance: A review,” *Energy Economics*, vol. 81, pp. 709–727, 2019, doi: https://doi.org/10.1016/j.eneco.2019.05.006.

[99] F. Rundo, F. Trenta, A. L. di Stallo, S. Battiato, “Machine learning for quantitative finance applications: A survey,” *Applied Sciences*, vol. 9, no. 24, p. 5574, 2019.

[100] Z. Lin, “Modelling and forecasting the stock market volatility of SSE Composite Index using GARCH models,” *Future Generation Computer Systems*, vol. 79, pp. 960–972, 2018.

[101] M. Frömmel, K. Lampaert, “Does frequency matter for intraday technical trading?,” *Finance Research Letters*, vol. 18, pp. 177–183, 2016.

[102] M. Schreyer, T. Sattarov, A. Gierbl, B. Reimer, Borth, “Learning sampling in financial statement audits using vector quantised variational autoencoder neural networks,” in *Proceedings of the First ACM International Conference on AI in Finance*, 2020, pp. 1– 8.

[103] R. I. Webb, D. D. Ryu, D. D. Ryu, J. Han, “The price impact of futures trades and their intraday seasonality,” *Emerging Markets Review*, vol. 26, pp. 80– 98, 2016, doi: 10.1016/j.ememar.2016.01.002.

[104] P. Bacchetta, E. Mertens, E. Van Wincoop, “Predictability in financial markets: What do survey expectations tell us?,” *Journal of International Money and Finance*, vol. 28, no. 3, pp. 406–426, 2009.

[105] L. Ryll, S. Seidens, “Evaluating the performance of machine learning algorithms in financial market forecasting: A comprehensive survey,” *arXiv preprint arXiv:1906.07786*, 2019.

[106] A. M. Rather, A. Agarwal, V. Sastry, "Recurrent neural network and a hybrid model for prediction of stock returns," *Expert Systems with Applications*, vol. 42, no. 6, pp. 3234–3241, 2015.

[107] J. P. Serbera, P. Paumard, "The fall of high-frequency trading: A survey of competition and profits," *Research in International Business and Finance*, vol. 36, pp. 271–287, 2016, doi: 10.1016/j.ribaf.2015.09.021.

[108] S. Alonso-Monsalve, A. L. Suárez-Cetrulo, A. Cervantes, D. Quintana, "Convolution on neural networks for high-frequency trend prediction of cryptocurrency exchange rates using technical indicators," *Expert Systems with Applications*, 1 2020.

[109] A. W. Li, G. S. Bastos, "Stock market forecasting using deep learning and technical analysis: a systematic review," *IEEE Access*, vol. 8, pp. 185232–185242, 2020.

[110] J. Patel, S. Shah, P. Thakkar, K. Kotecha, "Predicting stock market index using fusion of machine learning techniques," *Expert Systems with Applications*, vol. 42, pp. 2162–2172, 3 2015.

[111] M. Pratama, S. G. Anavatti, P. P. Angelov, E. Lughofer, "PANFIS: A novel incremental learning machine," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, pp. 55–68, 1 2014.

[112] G. Widmer, M. Kubat, "Learning in the presence of concept drift and hidden contexts," *Machine Learning*, vol. 23, no. 1, pp. 69–101, 1996, doi: 10.1007/BF00116900.

[113] D. Sahoo, Q. Pham, J. Lu, S. C. Hoi, "Online deep learning: Learning deep neural networks on the fly," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence IJCAI 2018*, Jul. 2018, pp. 2660–2666.

[114] A. L. D. Rossi, B. F. De Souza, C. Soares, A. de Leon Ferreira de Carvalho, C. Ponce, "A guidance of data stream characterization for meta-learning," *Intelligent Data Analysis*, vol. 21, no. 4, pp. 1015–1035, 2017.

[115] C. Käding, E. Rodner, A. Freytag, J. Denzler, "Fine- tuning deep neural networks in continuous learning scenarios," in *Computer Vision – ACCV 2016 Workshops*, Cham, 2017, pp. 588–605, Springer International Publishing.

[116] P. P. Angelov, *Empirical Approach to Machine Learning*. Springer, 2017.

[117] B. Krawczyk, L. L. Minku, J. Gama, J. Stefanowski, M. Woźniak, M. Wó Zniak, "Ensemble learning for data stream analysis: A survey," *Information Fusion*, vol. 37, pp. 132–156, 2017.

[118] R. D. Baruah, P. Angelov, "Evolving fuzzy systems for data streams: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, pp. 461–476, 11 2011.

[119] M. Pratama, S. G. Anavatti, M. Joo, E. D. Lughofer, "pclass: An effective classifier for streaming examples," *IEEE Transactions on Fuzzy Systems*, vol. 23, pp. 369–386, Apr. 2015, doi: 10.1109/TFUZZ.2014.2312983.

[120] M. Á. Abad, J. B. Gomes, E. Menasalvas, "Predicting recurring concepts on data-streams by means of a meta-model and a fuzzy similarity function," *Expert Systems With Applications*, vol. 46, pp. 87–105, 2015, doi: 10.1016/j.eswa.2015.10.022.

[121] J. Read, "Concept-drifting data streams are time series; the case for continuous adaptation," *arXiv preprint arXiv:1810.02266*, 2018.

[122] H. M. Gomes, J. Read, A. Bifet, J. P. Barddal, J. Gama, "Machine learning for streaming data: state of the art, challenges, and opportunities," *ACM SIGKDD Explorations Newsletter*, vol. 21, no. 2, pp. 6–22, 2019, doi: 10.1145/3373464.3373470.

[123] X. Zheng, P. Li, X. Hu, K. Yu, "Semi-supervised classification on data streams with recurring concept drift and concept evolution," *Knowledge-Based Systems*, vol. 215, p. 106749, 2021.

[124] M. Pratama, J. Lu, E. Lughofer, G. Zhang, S. Anavatti, "Scaffolding type-2 classifier for incremental learning under concept drifts," *Neurocomputing*, vol. 191, pp. 304–329, 2016.

[125] E. Lughofer, P. Angelov, "Handling drifts and shifts in on-line data streams with evolving fuzzy systems," *Appl. Soft Comput.*, vol. 11, pp. 2057–2068, Mar. 2011, doi: 10.1016/j.asoc.2010.07.003.

[126] Ł. Korycki, B. Krawczyk, "Streaming decision trees for lifelong learning," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2021, pp. 502–518, Springer.

[127] S. U. Din, J. Shao, "Exploiting evolving micro-clusters for data stream classification with emerging class detection," *Information Sciences*, vol. 507, pp. 404–420, 2020.

[128] R. Szadkowski, J. Drchal, J. Faigl, "Continually trained life-long classification," *Neural Computing and Applications*, pp. 1–18, 2021.

[129] N. Kasabov, D. Filev, "Evolving intelligent systems: Methods, learning, applications," in *2006 International Symposium on Evolving Fuzzy Systems*, Sept 2006, pp. 8– 18.

[130] S.-s. Zhang, J.-w. Liu, X. Zuo, "Adaptive online incremental learning for evolving data streams," *Applied Soft Computing*, vol. 105, p. 107255, 2021.

[131] M. Mermillod, A. Bugaiska, P. Bonin, "The stability- plasticity dilemma: Investigating the continuum from catastrophic forgetting to age-limited learning effects," *Frontiers in psychology*, vol. 4, p. 504, 2013.

[132] E. Lughofer, *Evolving fuzzy systems-methodologies, advanced concepts and applications*, vol. 53. Springer, 2011.

[133] M. Carnein, H. Trautmann, "Optimizing data stream representation: An extensive survey on stream clustering algorithms," *Business & Information Systems Engineering*, vol. 61, no. 3, pp. 277–297, 2019.

[134] A. Bifet, R. Gavaldà, G. Holmes, B. Pfahringer, *Machine learning for data streams: with practical examples in MOA*. MIT press, 2018.

[135] K. Namitha, G. Santhosh Kumar, "Learning in the presence of concept recurrence in data stream clustering," *Journal of Big Data*, vol. 7, p. 75, Dec. 2020, doi: 10.1186/s40537-020-00354-1.

[136] J. Gama, R. Sebastiao, P. P. Rodrigues, "On evaluating stream learning algorithms," *Machine learning*, vol. 90, no. 3, pp. 317–346, 2013.

[137] L. Yang, S. McClean, M. Donnelly, K. Burke, K. Khan, "Detecting and responding to concept drift in business processes," *Algorithms*, vol. 15, no. 5, 2022, doi: 10.3390/a15050174.

[138] G. Sateesh Babu, S. Suresh, G.-B. Huang, "Meta- cognitive Neural Network for classification problems in a sequential learning framework," *Neurocomputing*, vol. 81, pp. 86–96, 2011.

[139] A. Maslov, M. Pechenizkiy, I. Žliobaite˙, T. Kärkkäinen, "Modelling recurrent events for improving online change detection," in *Proceedings of the 2016 SIAM International Conference on Data Mining*, 2016, pp. 549–557, SIAM.

[140] P. P. Angelov, D. P. Filev, "An approach to online identification of takagi-sugeno fuzzy models," *Trans. Sys. Man Cyber. Part B*, vol. 34, pp. 484–498, Feb. 2004, doi: 10.1109/TSMCB.2003.817053.

[141] C. W. Chiu, L. L. Minku, "A diversity framework for dealing with multiple types of concept drift based on clustering in the model space," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[142] A. L. Suárez-Cetrulo, A. Cervantes, "An online classification algorithm for large scale data streams: igngsvm," *Neurocomputing*, vol. 262, pp. 67–76, 2017, doi: https://doi.org/10.1016/j.neucom.2016.12.093.

[143] E. Lughofer, C. Cernuda, S. Kindermann, M. Pratama, "Generalized smart evolving fuzzy systems," *Evolving Systems*, vol. 6, no. 4, pp. 269–292, 2015, doi: 10.1007/s12530-015-9132-6.

[144] P. E. Tsinaslanidis, D. Kugiumtzis, "A prediction scheme using perceptually important points and dynamic time warping," *Expert Systems with Applications*, vol. 41, no. 15, pp. 6848–6860, 2014, doi: https://doi.org/10.1016/j.eswa.2014.04.028.

[145] C. Martin, D. Quintana, P. Isasi, "Dynamic generation of investment recommendations using grammatical evolution.," *International Journal of Interactive Multimedia & Artificial Intelligence*, vol. 6, no. 6, 2021.

[146] B. Grün, "Model-Based Clustering," *Handbook of Mixture Analysis*, pp. 157–192, Feb. 2019, doi: 10.1201/9780429055911-8.

[147] P. D. McNicholas, "Model-based clustering," *Journal of Classification*, vol. 33, no. 3, pp. 331–373, 2016.

[148] F. Dellaert, "The expectation maximization algorithm," Georgia Institute of Technology, 2002.

[149] J. Vanschoren, J. N. Van Rijn, B. Bischl, L. Torgo, "Openml: networked science in machine learning," *ACM SIGKDD Explorations Newsletter*, vol. 15, no. 2, pp. 49–60, 2014.

[150] P. P. Angelov, X. Zhou, "Evolving fuzzy-rule-based classifiers from data streams," *IEEE Transactions on Fuzzy Systems*, vol. 16, pp. 1462–1475, Dec. 2008, doi: 10.1109/TFUZZ.2008.925904.

[151] O. Cordón, F. Gomide, F. Herrera, F. Hoffmann, L. Magdalena, "Ten years of genetic fuzzy systems: Current framework and new trends," in *Fuzzy Sets and Systems*, vol. 141, 2004, pp. 5–31.

Andrés L. Suárez-Cetrulo

Andrés received his BSc and MSc in Computer Science at Carlos III of Madrid (Spain) in 2013 and 2014, respectively. He received his PhD in Computer Science at University Carlos III of Madrid in 2022. He is currently a Data Science Architect at Ireland's National Centre for Applied AI, based at University College Dublin. His current interests focus on online machine learning for data streams, regime changes in financial markets, deep learning, transformers and generative models.

David Quintana

David Quintana holds Bachelor's degrees in Business Administration and Computer Science. He has an M.S. in Intelligent Systems from Universidad Carlos III de Madrid and a PhD in Finance from Universidad Pontificia Comillas (ICADE). He is currently Associate Professor at the Computer Science Department at University Carlos III of Madrid. There, he is part of the bio-inspired algorithms group EVANNAI. His current research interests are mainly focused on applications of Computational Intelligence in finance and economics.

Alejandro Cervantes

Graduated as Telecommunications Engineer at Universidad Politecnica of Madrid (Spain), in 1993. He received his PhD in Computer Science at University Carlos III of Madrid in 2007. He is currently a professor at the Escuela Superior de Ingeniería y Tecnología in UNIR (Universidad Internacional de la Rioja). His interests include bio-inspired algorithms for classification of non-stationary data, large multi-objective optimization problems, swarm intelligence algorithms and deep machine learning for meteorological forecasting, aeronautics and astrophysics.

An Improved Deep Learning Model for Electricity Price Forecasting

Rashed Iqbal[1], Hazlie Mokhlis[1], Anis Salwa Mohd Khairuddin[1,3]*, Munir Azam Muhammad[2]

[1] Department of Electrical Engineering, Faculty of Engineering, University Malaya, Kuala Lumpur (Malaysia)
[2] Department of Electrical Engineering, Main Campus, Iqra University, Karachi (Pakistan)
[3] Centre of Intelligent Systems for Emerging Technology, Faculty of Engineering, University Malaya, Kuala Lumpur (Malaysia)

* Corresponding author: anissalwa@um.edu.my

Received 6 January 2022 | Accepted 12 April 2023 | Published 6 June 2023

Abstract

Accurate electricity price forecasting (EPF) is important for the purpose of bidding strategies and minimizing the risk for market participants in the competitive electricity market. Besides that, EPF becomes critically important for effective planning and efficient operation of a power system due to deregulation of electricity industry. However, accurate EPF is very challenging due to complex nonlinearity in the time series-based electricity prices. Hence, this work proposed two-fold contributions which are (1) effective time series pre-processing module to ensure feasible time-series data is fitted in the deep learning model, and (2) an improved long short-term memory (LSTM) model by incorporating linear scaled hyperbolic tangent (LiSHT) layer in the EPF. In this work, the time series pre-processing module adopted linear trend of the correlated features of electricity price series and the time series are tested by using Augmented Dickey Fuller (ADF) test method. In addition, the time series are transformed using boxcox transformation method in order to satisfy the stationarity property. Then, an improved LSTM prediction module is proposed to forecast electricity prices where LiSHT layer is adopted to optimize the parameters of the heterogeneous LSTM. This study is performed using the Australian electricity market price, load and renewable energy supply data. The experimental results obtained show that the proposed EPF framework performed better compared to previous techniques.

Keywords

Intelligent System, LSTM, Smart Grid, Time Series, Forecasting.

DOI: 10.9781/ijimai.2023.06.001

I. Introduction

Smart grids (SG) are introduced to improve the performance of the traditional grid. With deregulation of electricity industry, electricity price forecasting (EPF) becomes critically important for effective planning and efficient operation of the power system. In several countries, deregulations of the electricity sector have been developed to enhance congestion control, facilitate renewable energy, and maximize the resource allocation of the power system [1]. Due to the significant volatility and intricate nonlinearity of electricity pricing, EPF has been a challenging issue. Accurate price forecasting has the ability to assist market participants to regulate their bidding strategies, production or consumption schedule with the intention to maximize their profits in the electricity market [2], [3]. Whenever demand is over- or under-predicted, inaccurate projections can have disastrous social and financial repercussions. Underestimating demand has a negative impact on supply, which leads to forced power interruptions and negative production effects. Meanwhile, overestimating demand may result in excessive investment in generation capacity, potential financial difficulty, and eventually increased electricity prices. Hence, this study plays an important role in the areas of power production and management with the aim to overcome the risk in electricity production investments and maintain affordable electricity price for the consumers [4].

Existing statistical techniques aim to reveal the specific pattern of historic power price by utilizing curve fitting. For instance, German electricity market has tested a k-factor Guégan Introduced Generalized Autoregressive Conditionally Heteroskedastic (GIGARCH) to forecast electricity price [5]-[6]. An iterative neural network methodology is also adopted along with this combinatorial neural network-based prediction technique to forecast upcoming electricity price. The advantages of this method include good precision, model functionality, and reliability. Meanwhile, Auto-regressive Integrated Moving Average (ARIMA) was proposed for short-term power load forecasting [7]. Application of statistical models had shown to be challenging when predicting multi-dimensional nonlinear price of electricity since they are mainly based on linear equations.

Please cite this article as:
R. Iqbal, H. Mokhlis, A. S. M. Khairuddin, M. A. Muhammad. An Improved Deep Learning Model for Electricity Price Forecasting, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 1, pp. 149-161, 2024, http://dx.doi.org/10.9781/ijimai.2023.06.001

On the other hand, shallow learning models have been proven to perform better compared to statistical models in terms of error minimization and some other factors. Due to nonlinearity and high volatility of the features in EPF, shallow learning models have shown to be feasible in electricity price forecasting [8]. In the field of load forecasting, Support Vector Machine (SVM) [9], [10] has been applied to predict ranges of nonlinear quantities and perform feature selection. Support vector regression (SVR) [11], artificial neural network (ANN) [18], [19], and regression tree are the main shallow machine learning models that have been commonly applied in forecasting system. Besides, the work in proposed a hybrid of SVR and gray wolf optimization to forecast life cost of power transformer. A hybrid model based on SVR and ANN is proposed in [16] by adopting new signal decomposition and correlation analysis technique to predict electricity price for next 24-hours. Furthermore, in [1] a hybrid approach of ANFIS and Backtracking Search algorithm (BSA) was proposed for electricity price forecasting and feature selection. Besides, a multi-objective binary-valued backtracking search algorithm (MOBBSA) and ANFIS approach have been employed. Nevertheless, over-fitting and gradient disappearance have been the common challenges in shallow machine learning models. It can be seen that previous techniques seemed to be less feasible for day-ahead EPF due to limited compatibility with big data and perplexing nonlinear problems [20]. The detail literature reviews related to this study is shown in the Table I.

Alternatively, deep learning algorithms have increasingly become popular in the disciplines of artificial intelligence and big data due to its ability to generate efficient classification approximations from a huge volume of input data and extract the data's underlying properties [21]-[23]. The model in [16] focused on distributed depiction, bidirectional gated recurrent unit (BiGRU) and learning algorithm with the BiGRU layer processing past and prospect information concurrently to fully extract chronological and nonstationary features from input data with the goal of improving forecasting performance. Meanwhile, to extract difficult nonlinear characteristics, [14] incorporated the deep belief network (DBN), LSTM RNN, and convolutional neural network (CNN). The DBN model was used in [24] to use signal processing and correlation analysis techniques. In addition, [25] created a multi-input and multi-output LSTM model for forecasting electricity demand. When evaluating the aerial correlation of dataset, it seems to be that a deep learning algorithm with a recurrent feedback framework called Recurrent neural network RNN has the capacity to accomplish more overarching and entire designing of time series than other traditional AI algorithms. The gradient inflation and gradient vanishing issues could be handled using LSTM through the RNN training procedure. As a result, LSTM has been used to anticipate day-ahead power prices for the Victoria region of Australia and the Singapore market [17]. Furthermore, the network topology of single gated recurrent units (GRU) has been explored for prediction purposes. When compared to an LSTM network, the GRU's simple neuron topology has been proven to lead to a faster processing time [26]. In a nutshell, LSTM has been demonstrated to perform better in terms of forecasting accuracy than SVM, ANN, and RNN [27], [28]. As a result, the analysis of time-series data for deep learning model in EPF has been an active subject of research for decades.

Based on the previous literatures, it can be seen that most of the works consider electricity supply, price and seasons to be the input features for the EPF system. Thus, in order to develop accurate prediction model, this work considers several inputs such as the price,

TABLE I. Comparisons of Recent Studies in Electricity Price Forecasting

Method	Application	RMSE	MAPE (%)	Limitations/Challenges
SVM[12] LSSVM[12]	Machine learning techniques are adopted to solve longer time horizon and highly nonlinear data for mid-term electricity market clearing price.	N/A	11.7491 10.9722	Accuracy in spike price forecasting considerably low by using the proposed machine learning methods. Optimization of forecasting accuracy in the spike price area is the main challenge of the study.
ANN PSO (Hybrid)[13]	Mid-Term Load Power Forecasting considering environment Emission using North American electricity market	N/A	1.9	ANN PSO method is not feasible to handle large data set of nonlinear data.
IFCM-SVM [9]	A dynamic parallel forecasting model using modified fuzzy time series and SVM. IFCM model is used to cluster the input data set, then the FTS model and SVM model are improved, finally the dynamic parallel model is used to forecast.	11.66	7.92	Computation of large data is a challenge. During forecasting process, the number of operation cycles need to be reduced in order to obtain good prediction accuracy.
k-factor GIGARCH [5]	This work combines several machine learning approaches to develop a novel hybrid forecasting model, namely EMD-SVR-PSO-AR-GARCH model. This work adopts the New South Wales (NSW, Australia) electricity market.	427 - 759	2.76 - 3.74	The nonlinearity and randomness of sequences of electricity consumption data.
GA-CNN [14]	The work is tested on Pennsylvania-New Jersey-Maryland (PJM) power market. The method integrates CNN with an evolutionary algorithm and utilizes spatiotemporal data.	0.007 – 0.02	3.5 – 4.9	Limited discussion on time series data analysis and statistical reliability.
EEMD-LSTM_ SMBO [15]	An optimized heterogeneous structure LSTM model is proposed to solve the problems of the single network structure and hyperparameter selection. PJM electricity market is adopted in this work.	0.9 – 1.9	2.5 – 4.7	Uncertain accuracy due to limited variables considered in the prediction model.
Bi-GRU (EGA-STLF) [16]	Bi-GRU layer in EGA-STLF computes the past and the future data simultaneously to fully extract temporal and nonlinear features from input data. This work adopts Australia electricity market for short-term load forecasting (STLF).	255.12	3.06	Analyze influence factors from more complex environments.
SCAR-Dvine model [17]	A flexible class of drawable vine copula models is applied by incorporating the dependence parameters of the constituting bivariate copulae to be time-varying. This work adopts Australia electricity market for the one-day-ahead forecasting.	N/A	N/A	Modelling risk of the five markets as a complex interconnected system, as opposed to analyzing markets individually.

demand, seasons, fuel supply, renewable and non-renewable energy, peak, and off-peak hours. Predicting the unknown source of spike can be a challenge in EPF. In light of this, an optimized improved deep learning framework is proposed by incorporating linear scaled hyperbolic tangent (LiSHT) layer in the EPF. The LiSHT layer is adopted to optimize the hyper parameters of the heterogeneous LSTM, to further improve the performance of the forecasting model and predicting the spikes.

Meanwhile, the unique properties and characteristics of time-series data are important for forecasting and prediction purposes. Time series data can be challenging due to presence of noise, exhibit high volatility and extremal directional movements [29]. Generally, stationarity of time series data is vital because various analytical tools and statistical models rely on it. Therefore, a pre-processing module is required to contribute towards accurate forecasting performance in terms of reliability and accuracy. In order to overcome this challenge, this work proposed a pre-processing module to ensure the feasibility of the time-series data to fit the proposed deep learning model. In the pre-processing module, a proper transformation is performed in order to satisfy the stationarity property of the time series data by applying Augmented Dickey–Fuller test and transformations. This is to prevent autocorrelation in the prediction model's errors. This will then contribute towards more accurate EPF. The contributions of this work are as follow:

1. Propose pre-processing module to monitor the suitability of the time series data for the deep learning model.
2. Propose an optimized RNN-LSTM based algorithm by incorporating LiSHT layer.

This paper is organized as follows. Section II discusses on the time-series analysis; Section III explains on the proposed forecasting model; Section IV presents the experimental results and discussion on case studies of the Australian electricity market. Finally, Section V concludes this study.

II. Data Pre-processing

A. Autocorrelation of the Model's Forecasting Reliability

In this research, the time series dataset includes electricity price, demand, and renewable energy supply of Australia's most important five economic zones. The electricity market data covers the duration from 1 September 2020 to 31 May 31 2021 which is obtained from https://aemo.com.au website. Conventional time series data may contain missing values, outliers and high dimensional data. These factors contribute to unstable forecasting performance. Therefore, pre-processing is required to solve the abovementioned problems. This work emphasized on linear trend-based equation for features processing. The linear trend approach can perform effectively with the trend and depict it without any assumptions. Besides, the residual seasonality, peak-off peak hour and renewable energy trend can distinguish any time series dataset shown in Table II.

Let h_1, h_2, … …, h_n be the time-series data. Equation (1) is the definition of a nonlinear regression model of order m is denoted by:

$$h_t = f(g_t, \theta) + \epsilon_t \tag{1}$$

where $g_t = (h_{t-1}, h_{t-2} \ldots .. h_{t-m}) \in \mathbb{R}^m$ made of m values of h_t, θ is the parametric vector and ϵ_t is the residual. After the model has been built, machine learning or deep learning approaches can be used to find the function $f(*)$. Root mean square error (RMSE) and mean absolute error (MAE) are the most often used indicators for regression performance evaluation of a forecasting model. Nevertheless, both regression performance evaluators only indicate the accuracy of the observed and estimated values. Since they are unable to analyse the fitness of time series data in the proposed forecasting model, the residuals are employed to assess this dedicatedly. In other words, the forecasting model's residuals of regression analysis for normal distribution and autocorrelation are estimated by function $\widehat{\epsilon_t}$, where $\widehat{h}_t$ is the predicted value as equation (2).

$$\widehat{\epsilon_t} = h_t - \widehat{h}_t \tag{2}$$

The presumption of no autocorrelation in the residuals might make the forecasting vulnerable as there may not be exploration on all available data in the training process. In other words, the reliance of the residuals indicates that the model did not well fit the time-series data and that there is important data remaining that must be investigated. Dataset used in this research showed in Table II. The autocorrelation function (ACF) plot and the Ljung–Box Q test for residual autocorrelation are two important techniques for -determining the presence of autocorrelation in the residuals [30]. More analytically, by calculating the linear correlation of every residual in various lags, $\hat{\epsilon}_{t-1}$, $\hat{\epsilon}_{t-2}$, … the ACF can be obtained in which the temporal autocorrelation is depicted by ACF, and the Ljung–Box Q test is a "portmanteau" test. The null hypothesis H_0 that ''a sequence of residuals does not exhibit autocorrelation for a specified number of lags L", is proved technically with respect to other hypothesis H1 that ''some autocorrelation coefficient is nonzero.'' Equation (3) defines the Ljung–Box Q test statistic in more detail,

$$Q = s(s+2) \sum_{k=1}^{M} \frac{\rho_k^2}{s-k'} \tag{3}$$

where equation (4) indicates at lag-k, autocorrelation coefficients ρ_k are,

$$\rho_k = \frac{\sum_{i=1}^{s-k} (h_i - \bar{h})(h_{i+k} - \bar{h})}{\sum_{i=1}^{s} (h_i - \bar{h})^2} \tag{4}$$

with $\bar{h} = \frac{1}{s} \sum_{i=1}^{s} h_i$ under H_0 the statistic Q asymptotically follows a $g_{(M)}^2$ distribution. The model shows autocorrelation and reject the zero hypothesis H_0 if as following equation (5),

$$Q > g_{(1-\alpha, M)}^2 \tag{5}$$

where the critical value of the Chi-square distribution is defined for significance level α, or critical level $p = 1 - \alpha$, known as p value.

TABLE II. Features Used in Time Series Data

Peak/ Off-Peak Hours	Time (Hour)	Main Electricity Supply (MWH)	Previous Hour Price (AUD)	Solar Power supply (MWH)	Hydro Power Supply (MWH)	Wind Power Supply (MWH)
1	5	7360.25	40.54	0	992	215.06
1	6	7066.01	43.59	0	992	192.84
1	7	6841.68	34.74	0	645.09	168.44
1	8	6732.33	17.15	0	325.58	146.70
1	9	6980.24	16.61	0	214.92	136.46
0	10	7661.34	31.56	0	125	186.70
0	11	8639.57	40.02	0	60	179.90
0	12	9890.74	49.99	0	2.5	810.19
0	13	9845.55	50.25	0	20	776.65
0	14	9446.45	54.32	0	94.58	819.08
0	15	8992.55	55.51	9.76	380.62	825.95
0	16	8547.70	47.99	189.73	705.43	760.85
0	17	8162.07	39.65	418.42	478.88	746.72
cont	cont	cont	cont	cont	cont	cont

B. Stationarity and No Stationarity

Autocorrelation, long memory, fractal and multi-fractal properties are the features of time-series that appear so frequently that they are referred to as stylized facts. The main disadvantage of working with values of price time series is that they follow a random walk process from the standpoint of stochastic processes. The coefficient of autocorrelation is ρ_k, with $k > 1$ are statistically remarkable for many lags L and the first-order autocorrelation coefficient ρ_1 is equal to one. This kind of time series are called unit root time-series or integrated of order one which are expressed by I (1). Modelling the levels of these series under such conditions is unproductive since the residuals of the models show redundancy, which putting the entire framework of statistical validity in jeopardy. In order to examine these series effectively, they must be stationary which is essential for the advent of a new forecasting model.

Assume that F_h $(h_{t_{1+\tau}}, \ldots\ldots, h_{t_{n+\tau}})$ is the total distribution algorithm of the intrinsic joint distribution of $\{h_t\}$ at times $t_1 + \tau, \ldots\ldots, t_1 + n$ then the stochastic process $\{h_t\}$ is strictly stationary if (6),

$$F_h\left(h_{t_1+\tau}, \ldots\ldots, h_{t_n+\tau}\right) = F_h\left(h_{t_1}, \ldots\ldots, h_{t_n}\right) \tag{6}$$

for all τ, $t_1 \ldots. t_n \in \mathbb{R}$ and $n \in N$. Nevertheless, the stationarity of time series is reduced resulting to weak covariance stationarity [30]. A stochastic process becomes covariance stationary when the mean is constant, the second moment is finite, and the covariance function relies on the difference between t_1 and t_2. Hence, in equation (7) the auto-covariance needs to be denoted with one variable, i.e.,

$$cov_{hh}(t_1, t_2) = cov_{hh}(t_1 - t_2, 0) \tag{7}$$

where cov_{hh} is the auto-covariance of the y_r series to summarize stationarity based on statistical features of the stochastic process. It has been a general hypothesis that many procedures such as statistical assessment, modelling and prediction become simpler when adopted the stationary processes. The partial autocorrelation function provides a resolution once the problem has been detected, where the lag-k coefficient $\phi_{k,k}$ is displayed by the indicated formula in equation (8),

$$\begin{cases} \phi_{k,k} = \dfrac{\rho_k - \sum_{j=1}^{k-1} \phi_{k-1,j}\rho_{k-j}}{1 - \sum_{j=1}^{k-1} \phi_{k-1,j}\rho_{k-j}} \\ \phi_{kj} = \phi_{k-1,j} - \phi_{k,k}\phi_{k-1,k-j}, \end{cases} \tag{8}$$

for $k > 1$ and $\phi_{1,1} = \rho_1$. Clearly, if there is unit root throughout the series, that is $\rho_1 = 1$, the first-order partial autocorrelation coefficient $\phi_{1,1}$ will become one. Commonly, the initial coefficient is statistically significant while the rest are insignificant. Then, the first series should be characterized by the first differences as sown in equation (9) of the series, defined by (9)

$$\triangle_t = h_t - h_{t-1} \tag{9}$$

Therefore, the first difference of the time series in stationarity obtained can be represented with integrated of order zero which is $I(0)$. However, crossover of different non-stationarities could present while computing the time-series data there such as unit-roots, structural pause, level up-downs, seasonal trend or a shifted variance. When the series is non-stationary (I (1)), the typical transformation is to take the first differences and transform it to stationary series (I (0)), whereas if the series contains structural breaks or a changing variance due to crises, a nonlinear BoxCox transformation will be the best solution [31]. As normality is an essential criterion for various statistical procedures, a BoxCox transformation provides a mechanism to turn non-normal data into a normal pattern. The following equation (10) defines one-parameter Box-Cox transformation as,

$$h_t = \begin{cases} \dfrac{h_t^{\lambda} - 1}{\lambda}, & \text{if } \lambda \neq 0 \ ; \\ \ln h_t, & \text{if } \lambda = 0. \end{cases} \tag{10}$$

where nonzero Box-Cox transformations are used for $\lambda =- 3,- 2,$ -0.5, 0, 0.5,1 and 2. The rule $\lambda = 0$ is followed by majority of the time series; therefore, the returns which are the first logarithmic differences are used to attain stationarity in these series as equation (11),

$$r_t = \ln h_t - \ln h_{t-1} \approx \frac{h_t - h_{t-1}}{h_{t-1}} \tag{11}$$

the last expression being the percentage change or returns [30].

C. Augmented Dickey Fuller (ADF) Test

The proposed pre-processing module for greatly improving the accuracy and durability of a deep learning algorithm for time series prediction is discussed in this part, based on well-known statistical concept and estimation for stationarity and non-stationarity qualities. Generally, the components of the dataset are not-stationary when a machine learning or deep learning model is applied to estimate the time-series. This implies that they may have unit roots and some order of integration. It is worth noting that the Augmented Dickey–Fuller (ADF) test has the ability to identify a unit root in a time series data [30], [32]. The model is subjected to the testing method as following equation (12),

$$\triangle_{h_t} = \alpha + \beta t + \gamma h_{t-1} + \sum_{i=1}^{k-1} \delta_i \triangle_{h_{t-i}} + \epsilon_t \tag{12}$$

where ρ_1 denotes the first-order autocorrelation coefficient and α is a constant, β is the coefficient of trend, and $\gamma = (\rho_1 - 1)$. It is notable that k is the lag order of the autoregressive determined so that the residuals ϵ_t have no serial correlation. There has a stochastic random walk process, if $\alpha = 0$ and $\beta = 0$, while if $\alpha \neq 0$ and $\beta = 0$, here the stochastic process is with drift. The unit root test is employed to evaluate statistical importance under the null hypothesis. H_0:$\{\gamma = 0$ that is $\rho = 1\}$ versus the nonzero hypothesis H_1: $\{\gamma < 0$ that is $\rho < 1\}$.

Recursively taking the first differences in (9) or returns in (11) until the sequence is made stationary depending on the nature of the series. The autocorrelation in the model's residuals will be reduced when using a series of transformations based on the first difference and returns. This means that the forecasting method will be considerably better at explaining the data because it captures all conceivable nonlinearities, assuring the model's accuracy and efficacy.

The pseudo-code for the framework is shown below pseudo-code. Firstly, the time-series data is imported. The ADF test is then used to determine if the sequence levels are non-stationary, or if they have a unit root in Step 2. If the sequence is stochastic, the dataset will be continuously converted using first differences or returns till the resultant series becomes stationary in Steps 4–7. The newly modified time-series data is then utilised to train the forecasting model in Step 8.

Pseudo-code of proposed algorithm

```
1. Input time series data
2. Assess unit root test (ADF)
 3. If (Unit root exist in time-series) then
 4. Repeat
  5. Convert the time series based on differences (9) or returns (11).
  6. Assess unit root test (ADF).
  7. Until (Stationarity exists in time-series.)
  8. By using converted time-series, train the prediction model.
 Else
    9. By using real time-series, train the prediction models.
    10. On the training sample, estimate the residuals.
    11. If (Autocorrelation exists in residuals.) then
    12. Convert the time series based on differences (9) or returns (11).
    13. Using the converted time-series, retrain the forecasting model.
       End
  End
```

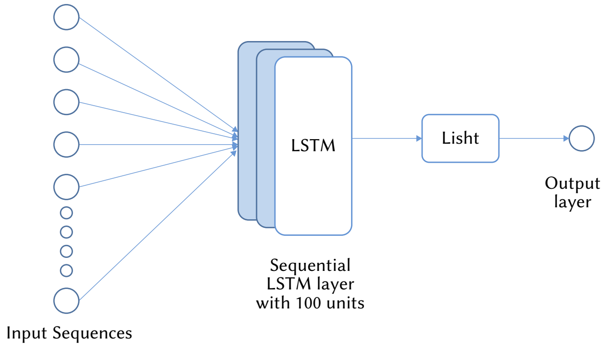


Fig. 1. LSTM+LiSHT prediction model architecture.

If the data is stationary, then levels of time-series are employed to train the forecasting model in Step 10. The errors of the estimation method on the learning algorithm are employed for further analysis and testing. It is noticeable that a training is performed with a series that has a unit root then the predicted values become near to the real values for any realistic model then, the presence of strong autocorrelation factors marks the model as unproductive [33]. Therefore, ACF plots and/or the Ljung–Box Q test are used to investigate autocorrelation within residuals of the dataset in step 11. Eventually, if the residuals have autocorrelation, the recommended transformation is performed to the training phase and the algorithm is retrained utilizing the newly transformed dataset according to steps 12–13. It is noticeable that if the series levels are stationary and the residuals on the training dataset indicate no autocorrelation, there is no need to reform the series because it will result in catastrophic phenomena of over-differencing. To put it in another way, over-differencing makes the entire mechanism "non-invertible," and thus lacked an endless autoregressive expression. In the form of a flowchart, Fig. 2 depicts an insight into the intended structure.

Finally, if the classifier is trained with a transformed series with no autocorrelation in residuals, the inverse transformation will be used in the model's forecasts to obtain the forecast for the levels of the exact time-series and reliable for parametric and no parametric tests.

III. The Proposed Improved LSTM Forecasting Model

In order to process the long sequence of time series data, LSTM Recurrent Neural Network (RNN) is proposed with the aim to overcome the problem of vanishing gradient and gradient explosion that can occur in conventional RNN. The input gates and output gates are replaced by memory/forget gates in the hidden layer of LSTM RNN which include memory space and information flow process for long historical time series [14]. A sequential layer followed by a fully connected layer, lstm layer, tanh layer and regression layer are applied in this algorithm Fig.1. In this study, sgdm optimization is applied with max epoch 1500. Gradient threshold is 1, learning rate schedule piecewise. In lstm layer, number of hidden units is 100, state activation function tanh and sigmoid gate activation function are adopted in this algorithm. The parameters of LSTM are shown in Table III.

TABLE III. Parameters Used in Lstm

LSTM Parameters	
Hidden layers	3
Neurons per layer	100
Type of layer	LSTM
Activation layers	Sigmoid
Epochs	1500
Optimizer	SGDM

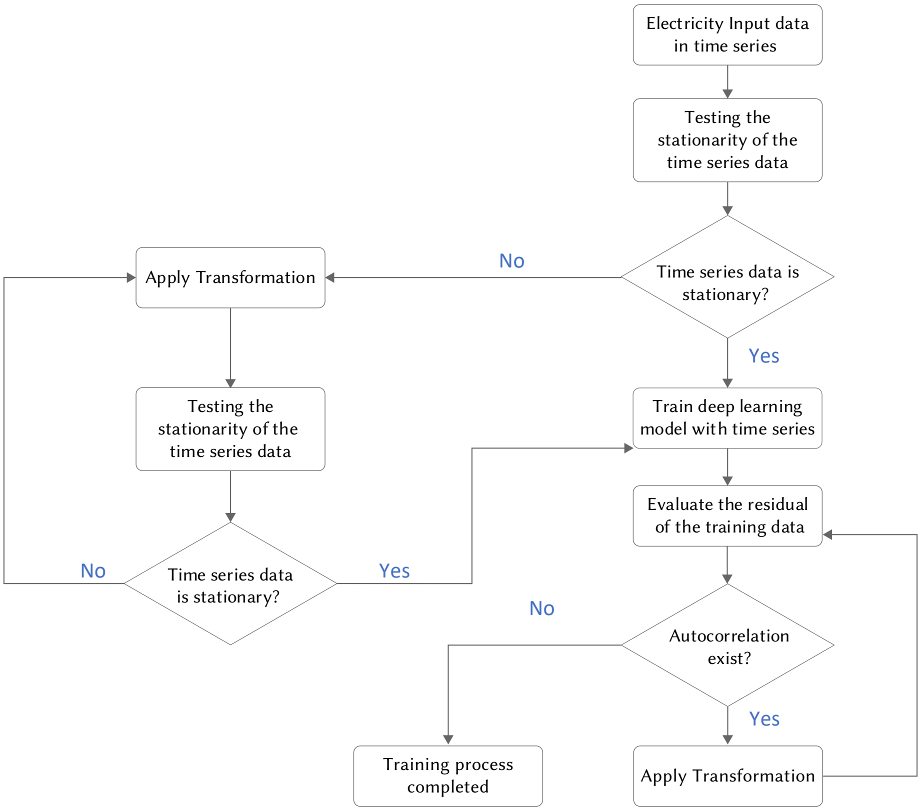

Electricity Input data in time series
Testing the stationarity of the time series data
Time series data is stationary?
No
Apply Transformation
Yes
Testing the stationarity of the time series data
Train deep learning model with time series
Evaluate the residual of the training data
No
Time series data is stationary?
Yes
No
Autocorrelation exist?
Yes
Training process completed
Apply Transformation

Fig.2. Flowchart of the proposed algorithm.

Conventional activation functions such as ReLU and Swish are less feasible for large negative input values and also may suffer from the dying gradient problem due to zero-hard rectification. Therefore, it is essential to adopt a better activation function to overcome those limitations. In this work, a non-parametric function, called Linearly Scaled Hyperbolic Tangent (LiSHT) for Neural Networks (NNs) is employed in this model as referred (13). The LiSHT activation function is utilized to scale the non-linear Hyperbolic Tangent (Tanh) function through a linear function and tackle the dying gradient problem.

Let an input vector be a ∈ R^d, and each hidden layer is capable to transform its input vector by applying a nonlinear mapping from the q^{th} layer to the $(q + 1)$ th layer as equation (13):

$$a = \tau^0$$
$$\sum_{l=1}^{N^q} w_{kl}^q \tau_l^q + o_k^q = c_k^{q+1}$$
$$\phi\left(c_k^{q+1}\right) = \tau_k^{q+1} \tag{13}$$

LiSHT is a non-parametric linearly scaled hyperbolic tangent activation layer that has unrestricted upper limits property on the right-hand side of the activation curve. LiSHT has the advantage of positive activation that does not identically propagate for all inputs, which solves the gradient problem at back propagation and contributes to faster training of the deep neural network. The LiSHT activation function is calculated by multiplying the *Tanh* function by its input *x* and defined as the equation (14) and (15). where $g(x)$ is a hyperbolic tangent function [32].

$$\phi(x) \; = \; x \cdot g(x) \tag{14}$$

$$g(x) = Tanh(x) = \frac{exp^x - exp^{-x}}{exp^x - exp^{-x}} \tag{15}$$

IV. Results and Discussions

In this work, the data were divided into training and test set consisting of hourly electricity price data as tabulated in Table IV:

TABLE IV. Seasonal Training Dataset

Season	Training set	
	1 day forecasting	1 month forecasting
Sep-Oct-Nov (Spring)	Oct (696 hours)	Sep-Oct (1440 hours)
Dec-Jan-Feb (Summer)	Jan (696 hours)	Dec-Jan (1440 hours)
Mar-Apr-May (Autumn)	Apr (696 hours)	Mar-Apr (1440 hours)

There have been no missing data in any of the time-series and outlier prices were not eliminated in order to preserve the characteristics of every series, even though these prices are the consequence of rare events presents the descriptive analysis for every training dataset and testing dataset, such as the measurements of minimal, max, average, sample variance (std. dev.), median, skewness, and kurtosis for illustrating the distribution's nature. Using the ADF unit root test, the proposed framework employed the National Electricity Market (NEM) price time-series in Australia to determine whether the training data are stationary or not. The outputs of the ADF unit root test for the training data of Australia's five states series under investigation are shown in Table V. Considering the t-statistics (t-stat) and the corresponding p values the null hypothesis H_0: ''the levels possess a unit root and are non-stationary'' is accepted for time series.

TABLE V. ADF Unit Root Test for the Training Data

Series	NSW	QLD	SA	TAS	VIC
***t* static**	-37.16	-71.36	-28.82	-39.42	-24.58
***p* value**	0.000*	0.000*	0.000*	0.000*	0.000*

In the sequel, the ADF test was run on a time-series to see if the unit root existed, as per the provided framework. The outcomes of the ADF unit root test for the training data of all time-series datasets are shown in Table IV. The (*) indicates statistical impact at the 5% critical threshold. Clearly, it's worth noting that all p values are almost zero, the null hypothesis H_0 is rejected.

As a result, the time series are "appropriate" for training a deep learning model with minimal autocorrelation in the errors, and a significant boost in forecasting accuracy is anticipated when comparing to same model trained with the non-transformed series. In order to evaluate the performance of the proposed model, the regression ability is assessed using mean absolute error (MAE) and root mean square error (RMSE). Besides that, another four key performance indicators are also employed: Accuracy (Acc), F1-score (F1), Sensitivity (Sen), Specificity (Spe), Positive Predicted Values (PPR) and Negative Predictive Values (NPV) which are indicated by the following equations (16)-(21).

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{FP}'} \tag{16}$$

$$\mathrm{F}_1 = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}'} \tag{17}$$

$$\mathrm{Sen} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}'} \tag{18}$$

$$\mathrm{Spe} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}'} \tag{19}$$

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}'} \tag{20}$$

$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}} \tag{21}$$

In this case, TP represents the frequency of prices that were successfully identified as raised, the number of prices that were successfully detected as having a decreasing value is denoted by TN, FP is the amount of prices that were incorrectly detected as being increased, whereas FN denotes the quantity of prices that were incorrectly detected as being dropped. Furthermore, the area under curve (AUC) statistic, considered one of the most important classification metrics which has been incorporated in the assessment and is shown using the receiver operating characteristic (ROC) curve. The ROC curve is made by comparing the true positive rate (Sensitivity) against the false positive rate (Specificity) at different cut-off values.

A. Pre-processing of Time Series Data for the Prediction Model

In the following section, the predictability of all prediction techniques is investigated by using the Auto-Correlation Function (ACF) plot and the Ljung–Box Q test to detect autocorrelation in the residuals. This is to ensure that each trained model adequately fits the time-series and if they are uniformly distributed evenly and monotonically independent. The Ljung– Box Q test is a ''portmanteau'' test which analyse the null hypothesis H_0 that ''a series of residuals exhibits no autocorrelation for a fixed number of lags L,'' which is the opposite of another hypothesis H_1 that ''some autocorrelation coefficient is nonzero coefficient is nonzero''. Fig. 3 – 7 displayed the ACF graphs of LSTM+LiSHT model for the electricity price data. The ACF graphs of the prediction model are trained with typical time-series defying the presumption of the free autocorrelation residuals. The high spikes which observed in several lags (fig. 3(a) – 7(a)), show that such model's estimation may be inaccurate. The ACF plot spikes at lag 1 then slowly decays to lag 10. From lag 1 to 4 spikes are too high and cut off at the significant band 0.2, Which shows that the significant autocorrelation presents in the residual of trained data. On the contrary, from the fig 3(b)-7(b) it is shown that the spikes from lag 2 immediately go down under or between the significant band. Therefore, the autocorrelation in the residual does not exist in the trained data and is statistically sound for the evaluation of time series. In summary, all ACF plots of the LSTM+LiSHT generated using the converted time series (figs 3(b) – 7(b)), show that the residuals do not have autocorrelation. This can be further verified by the results obtained from the Ljung– Box Q test

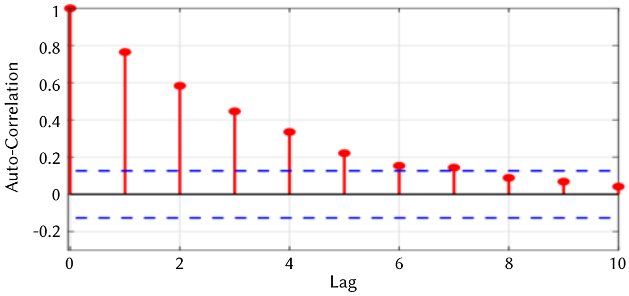


Fig. 3(a). Autocorrelation of residuals for NSW time series.

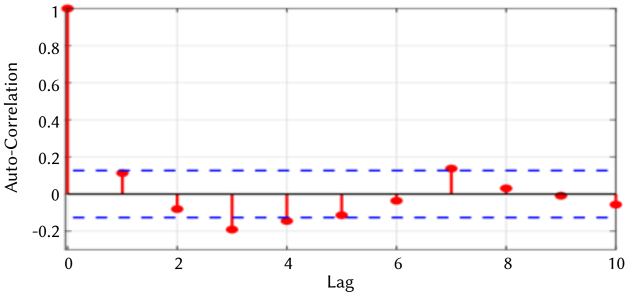


Fig. 3(b). Autocorrelation of residuals for transformed (box-cox) NSW time series.

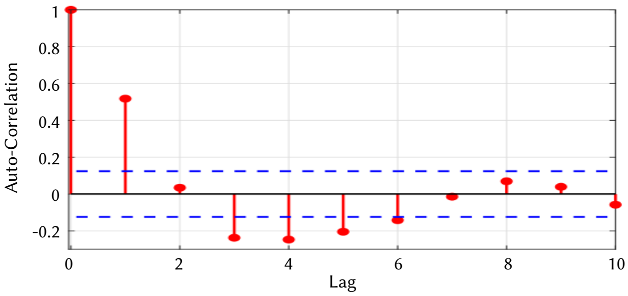


Fig. 4(a). Autocorrelation of residuals for QLD time series.

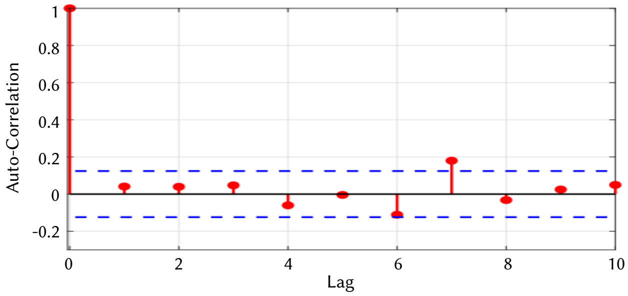


Fig. 4(b). Autocorrelation of residuals for transformed (box-cox) QLD time series.

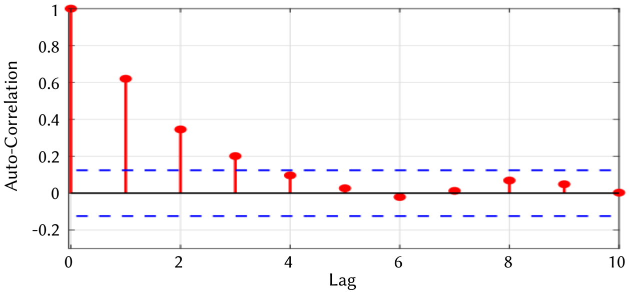


Fig. 5(a). Autocorrelation of residuals for SA time series.

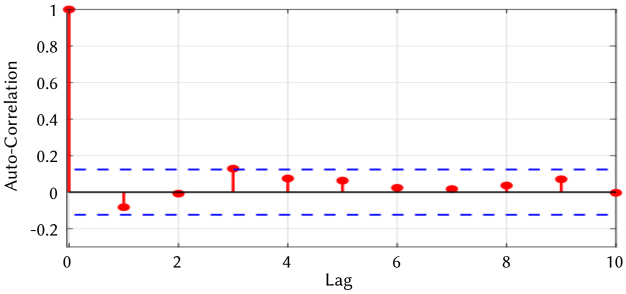


Fig. 5(b). Autocorrelation of residuals for transformed (box-cox) SA time series.

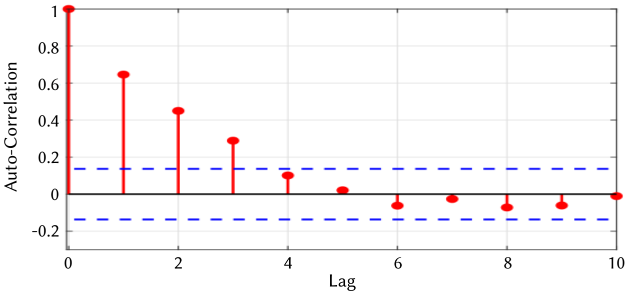


Fig. 6(a). Autocorrelation of residuals for TAS time series.

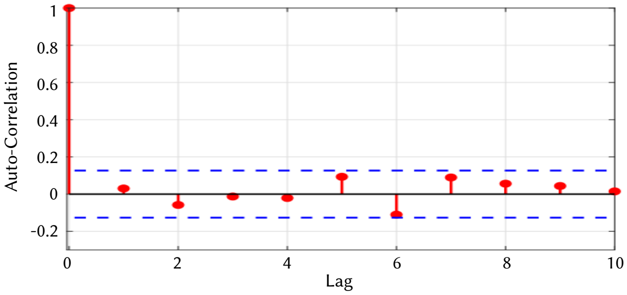


Fig. 6(b). Autocorrelation of residuals for transformed (box-cox) TAS time series.

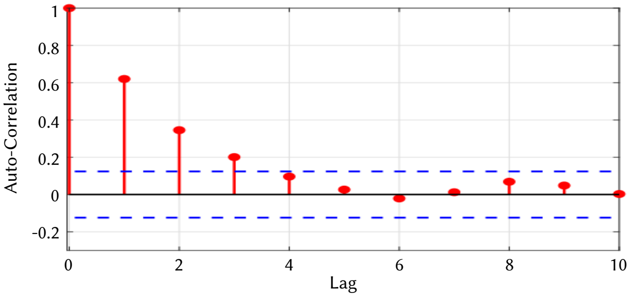


Fig. 7(a). Autocorrelation of residuals for VIC time series.

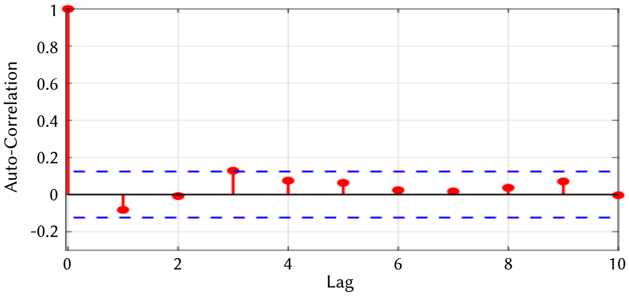


Fig. 7(b). Autocorrelation of residuals for transformed (box-cox) VIC time series.

(Table VI) where the transformed time series data using the BoxCox transformation is free from autocorrelation.

TABLE VI. Represent the Result of the Ljung–Box Q test Using L = 10

Condition	Forecasting	Autocorrelation existence
Before transformed	1 day	Yes
	1 month	Yes
After transformed	1 day	No
	1 month	No

In this work, it has been established theoretically and experimentally that the time-series data are ‘‘appropriate’’ for developing a deep learning model, which is one of the contributions of this study. In another way it can be said that this work has developed a new framework that can discover effective time series data for training a deep learning model. This will lead to a stable and reliable forecasting model. The term “appropriate” denotes that the time-series data has satisfied the stated scientific requirements and is adequate for training a forecasting model. If, on the other hand, the series fails to meet the desired criteria, it is deemed “unsuitable,” and any attempts to develop a solid prediction model would most likely be useless. Therefore, this work is a beginning point for the development of any prediction methodology for various time series forecasting. If the starting dataset is unstable or non-stationary, the work done for developing the forecasting model could be meaningless. Furthermore, it can be justified that this work has developed an innovative and comprehensive framework that allows any unstable time-series to be transformed to a stable condition by conducting a boxcox transformation method. It can be seen that the proposed transformation has successfully eliminated the “unsuitable” data, avoiding the costly and time-consuming “trial and error” method. Besides that, it is noticeable that one of the most interesting properties of our suggested framework is that this method can be simply modified to encompass a broader scientific domain of time-series forecasting operations without requiring any further adjustments or limits. More specifically, the recommended method uses statistic and economic tests to conduct an optimal pre-processing phase for utilising the internal structure of the timeseries. Finally, it is seen that while deep learning models are well accepted for time series, the proposed framework significantly enhances performance. However, more study is being done to see which of these approaches may be implemented more effectively a priori based on the properties for every time-series in order to get better forecasting performance. For accomplishing the prerequisite diagnosis and appropriate time transformation methodology, a complex pre-processing framework that refers to the inherent time-series particular traits such as stationarity, heteroskedasticity, seasonal cycles, and shifting variance can be used.

B. Forecasting Performance of the Proposed LSTM+LiSHT Model

The efficacy of the proposed LSTM+LiSHT prediction model for the energy price dataset during spring season is presented in Table VI, while the results for other seasons are presented in Appendix A. In spring, the accuracy of the proposed LSTM+LiSHT model is above 0.95 and 0.87 for the case of one day forecasting and one month forecasting respectively. Commonly, Medium-Term Forecast (MTF) studies horizons from a few days to months ahead. MTF is normally used for risk management, balance sheet calculations, and derivatives pricing. Meanwhile, Short-Term Forecast (STF) covers horizons from a few minutes up to a few days ahead has become an essential tool for the daily market operations.

Table VII also computed the sensitivity and specificity of the proposed forecasting model for both forecasting horizons. The sensitivity analysis of the proposed model is computed to permit the analysis of changes in expectations used in forecasting the electricity price. By studying all the variables and the possible outcomes, important decisions can be made about businesses, the economy, and making investments. On the other hand, high specificity indicates the good capability of the proposed model to avoid false alarms or false spikes in forecasting the electricity prices. The sensitivity and specificity of the proposed forecasting model is considered good performance which is above 0.7 and 0.8 respectively for both 1 day forecasting and 1 month forecasting. As such, sensitivity and specificity analysis are very useful methods to be applied in investment appraisal, sales and profit forecasting and other quantitative aspects of business management.

Table VIII presented the performance of the forecasting model without applying the transformation method in the pre-processing module. More specifically, the proposed model had shown to be biased when it was trained using the conventional time-series data which resulted to low forecasting performance. The forecasting accuracies are in the range between 0.4 to 0.8 and 0.3 to 0.8 for 1 day forecasting and 1 month forecasting respectively. The sensitivity and specificity are significantly low as well. The forecasting sensitivity is below 0.6 for both 1 day forecasting and 1 month forecasting. The sensitivity and specificity are significantly low as well. Hence, it is important to adopt the proposed pre-processing module to transform the conventional time series data in order to improve the forecasting performance.

As compared to table VII, the proposed forecasting model exhibits better performance in terms of ACC, sensitivity, and specificity when the data is trained with the transformed time series in the proposed pre-processing module. Furthermore, the interrelation between sensitivity and specificity has been significantly improved. In summary, the performance of the proposed LSTM+LiSHT forecasting model has been considerably improved after adopting the first transformed box-cox time series data, instead of the conventional time-series data. This justifies the contribution of the proposed pre-processing module in this work.

TABLE VII. Performance of the Proposed LSTM+LiSHT Model for Spring Season (After Transformation)

Australia’s region	Forecasting	ACC	AUC	F1	Sen	Spe	PPV	NPV
NSW	1 day	1	0.945	1	1	1	1	1
	1 month	0.984	0.802	0.852	1	0.983	0.742	1
QLD	1 day	1	0.957	1	1	1	1	1
	1 month	0.870	0.768	0.678	0.76	0.894	0.612	0.944
SA	1 day	0.958	0.896	0.909	1	0.947	0.833	1
	1 month	0.923	0.853	0.927	1	0.851	0.864	1
TAS	1 day	0.958	0.965	0.933	0.875	1	1	0.941
	1 month	0.970	0.910	0.977	0.997	0.921	0.958	0.995
VIC	1 day	1	0.962	1	1	1	1	1
	1 month	0.997	0.955	0.9583	1	0.997	0.92	1

TABLE VIII. Performance of the Proposed LSTM+LiSHT Model for Spring Season (Before Transformation)

Australia's region	Forecasting	ACC	Sen	Spe
NSW	1 day	0.583	0	0.636
	1 month	0.885	0.090	0.936
QLD	1 day	0.800	0.500	0.826
	1 month	0.761	0.183	0.818
SA	1 day	0.400	0.500	0.380
	1 month	0.332	0.574	0.226
TAS	1 day	0.680	0.500	0.695
	1 month	0.785	0.333	0.825
VIC	1 day	0.800	0.500	0.826
	1 month	0.850	0.065	0.976

TABLE IX. One Day Forecasting for New South Wales

Seasonality	Error	BILSTM	LSTM+GRU	GRU	LSTM	The proposed work
Spring	RMSE	2.7418	2.1190	1.5455	1.3222	4.197x10-6
	MAE	2.0013	1.6080	1.0110	0.8655	0.023
	MBE	0.6502	0.2198	0.1456	0.0443	0.0067
	MSE	7.5173	4.4903	2.3884	1.7481	0.0004
	R	0.9998	0.9999	0.9984	0.9993	0.9998
Summer	RMSE	4.3231	0.8878	0.8254	0.6725	2.598x10-6
	MAE	4.1249	0.7626	0.6944	0.4768	0.0094
	MBE	4.1249	0.5250	0.3768	0.2045	0.1822
	MSE	18.6895	0.7882	0.6814	0.4523	0.0003
	R	0.9967	0.9988	0.9992	0.9992	0.9997
Autumn	RMSE	3.1852	0.4122	0.2081	0.2430	1.730x10-6
	MAE	2.8768	0.3114	0.1596	0.1762	0.0018
	MBE	2.8550	-0.1127	0.0659	0.0427	0.0599
	MSE	10.1452	0.1699	0.0433	0.0590	0.0009
	R	0.9972	0.9992	0.9998	0.9997	0.9999

TABLE X. One Month Forecasting for New South Wales

Seasonality	Error	BILSTM	LSTM+GRU	GRU	LSTM	The proposed (LSTM+LiSHT)
Spring	RMSE	3.5388	0.5678	0.5630	0.4507	0.4065
	MAE	3.1164	0.3811	0.3236	0.2492	0.2530
	MBE	2.6034	0.0205	0.0192	0.0352	0.0134
	MSE	12.5229	0.3224	0.3170	0.2032	0.1652
	R	0.9991	0.9997	0.9997	0.9997	0.9998
Summer	RMSE	3.5682	0.7065	0.7134	0.5407	0.4207
	MAE	3.3058	0.3294	0.3274	0.2634	0.2329
	MBE	3.2019	0.0254	0.0498	8.6863e-04	0.0255
	MSE	12.7318	0.4991	0.5090	0.2924	0.2285
	R	0.9956	0.9964	0.9978	0.9978	0.9984
Autumn	RMSE	16.2570	7.6945	7.5842	5.5981	5.4047
	MAE	14.4163	2.6332	2.2764	2.3837	2.4569
	MBE	13.0280	0.1859	0.9894	0.5336	0.3663
	MSE	264.290	59.2051	57.520	31.3386	29.2107
	R	0.9983	0.9980	0.9980	0.9989	0.9990

In addition, the proposed forecasting model is compared with several deep learning models as tabulated in Table IX and Table X for one day forecasting and one month forecasting respectively. More comparison results which cover other regions and seasons are presented in the Appendix. Besides that, the graphical representation of the actual electricity price and predicted electricity price for one day (24 hours) forecasting and 1 month forecasting for five different states over the spring season are presented in Figs. 8 (a-e) and 9 (a-e), respectively. The regression analysis has been performed to quantify the relationship between variables used in the forecasting models. As shown in Table VIII and Table IX, the regression (R) values for all the forecasting methods are approximately 1.0 which indicates a good relationship between the forecast variable of interest and the predictor variables. Generally, RMSE which represents the standard deviation of residuals (forecasting errors), is computed to determine the concentration of data around the line of best fit. Meanwhile, MAE computed the average of all differences between actual and forecast absolute value. Another common metric applied is the mean

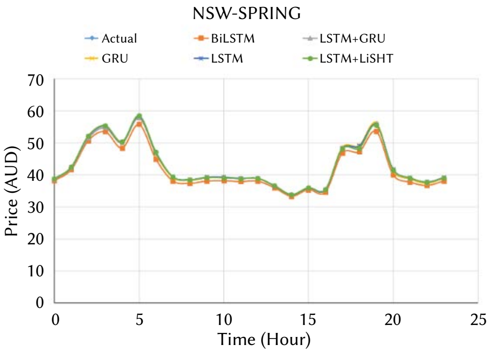


Fig. 8(a). One day prediction results for NSW.

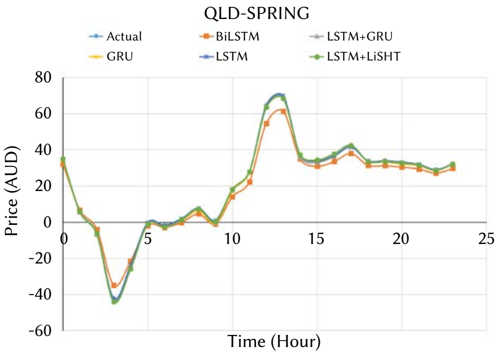


Fig. 8(b). One day prediction results for QLD.

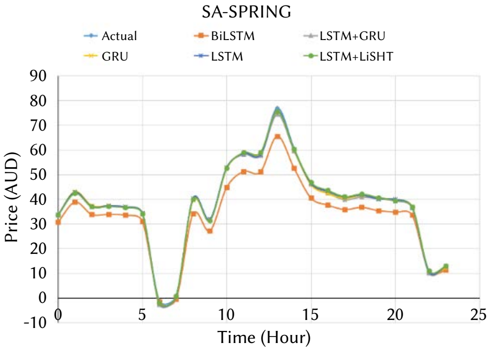


Fig. 8(c). One day prediction results for SA.

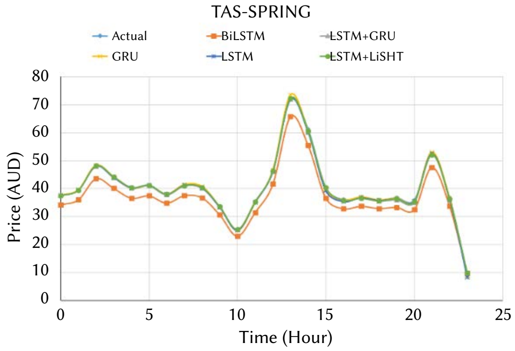


Fig. 8(d). One day prediction results for TAS.

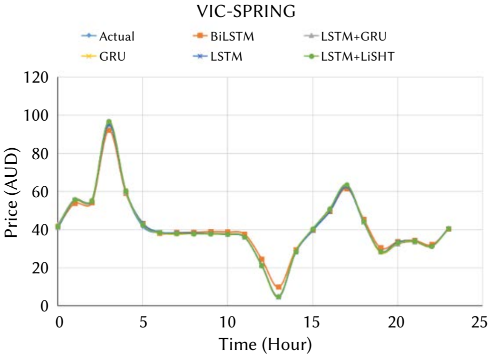


Fig. 8(e). One day prediction results for VIC.

bias error (MBE) which could provide indication whether the model overestimates or underestimates the output. The smaller the values of RMSE, MBE and MAE indicate better performance of the forecasting model. It can be seen that the lowest RMSE, MBE and MAE are achieved by the proposed LSTM+LiSHT forecasting model when compared to other forecasting models. In contrast, BiLSTM forecasting model produces the highest RMSE, MBE and MAE for spring, summer and autumn respectively. This shows that BiLSTM is the least preferable forecasting model to be used in this work followed by LSTM+GRU, GRU and LSTM. Hence, this can be justified from figs 8 (a-e) and 9 (a-e), where the curves of the proposed model coincide with the curve of the actual data which shows that the proposed model is able to forecast the electricity price effectively unlike the BiLSTM curves.

The proposed model is benchmarked with previous works as tabulated in Table XI. The work in [34] shows that the proposed Bi-GRU and Gated-FCN obtains RMSE of 8.23 and 3.12 respectively. Besides, the work in [35] that applied CNN-LSTM obtains RMSE of 6. The work in [36] applied BP, CNN, LSTM-NN, WT-TDLSTM model for electricity price forecasting and achieved the considerably low RMSE of 0.012798, 4.697257 × 10^{-5}, 0.008360, and 3.940309 × 10^{-6} respectively. Moreover, the work in [36] shows the WT_TDLSTM model is superior compared to the neural network system that exclude discrete wavelet transform and pre-processing of multiscale data. As for this work, the proposed model is tested on a smaller dataset that spans from year 2020 to 2021 and has produced considerably low

TABLE XI. Performance Comparison of the Proposed EPF Model With Recent Works

Work	MSE	RMSE	MAE
Bi-GRU [34]	N/A	8.23	N/A
CNN-LSTM [35]	N/A	5.92	N/A
Gated-FCN [34]	N/A	3.12	N/A
BP [36]	0.113129	0.012798	0.281345
CNN [36]	0.006854	4.697257 × 10^{-5}	0.071783
LSTM-NN [36]	0.091435	0.008360	0.22594
WT_TDLSTM [36]	0.001985	3.940309 × 10-6	0.035024
The proposed work	0.00053	2.8438× 10-6	0.0114

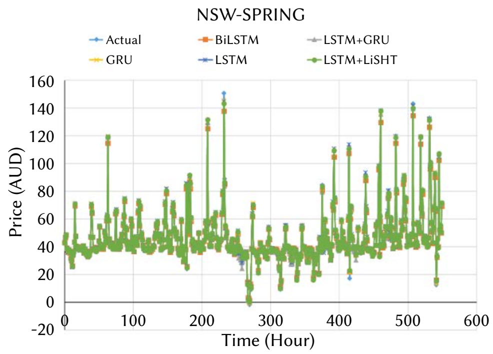


Fig. 9(a). Monthly prediction results for NSW.

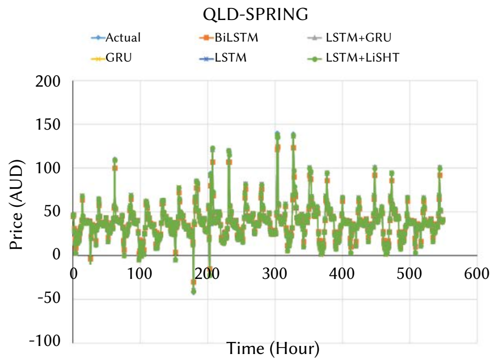


Fig. 9(b). Monthly prediction results for QLD.

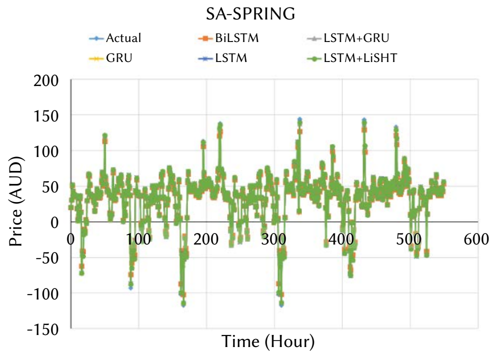


Fig. 9(c). Monthly prediction results for SA.

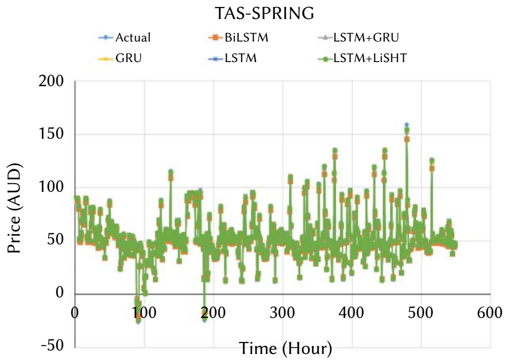


Fig. 9(d). Monthly prediction results for TAS.

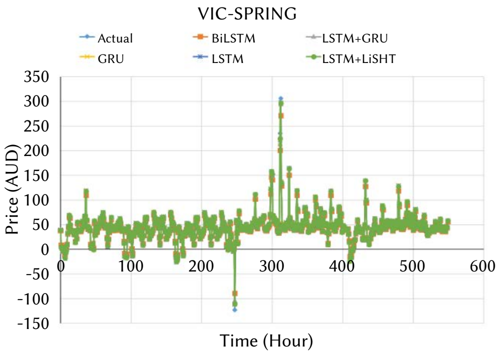


Fig. 9(e). Monthly prediction results for VIC.

RMSE, MAE and MSE as compared to previous works. The RMSE of the proposed LSTM+LiSHT framework varies for different season and region with the lowest value of 2.8438 × 10^{-6}. This justifies that the proposed forecasting model is suitable to be applied in the EPF application under various seasons in Australian electricity market.

V. Conclusion

In this work, time series data analysis has been performed and improved deep learning method has been proposed for short term electricity price forecasting. The developed forecasting model consists of pre-processed and post trained data analysis which incorporates time series statistical reliability method. An augmented dickey fuller test is performed to examine the stationarity and nonstationary data before the training process. Then, autocorrelation of the residuals is computed after the training process. In this work, the autocorrelation of the residuals has been evaluated to ensure the feasibility of the data for the deep learning approach. The autocorrelation in residuals has been fixed by transforming the data through box-cox transformation technique. Finally, the forecasting of electricity price is performed by applying the proposed deep learning module which has been modified to optimize the parameters of the heterogeneous LSTM. The performance of the proposed forecasting model has been benchmarked with previous works to justify the feasibility of the proposed method. Based on the results obtained, it can be seen that the proposed model has shown superior results compared to other methods in terms of RMSE, MSE and MAE.

In future works, further analysis can be performed such as comparing the proposed method for new profit and return-based performance measurements. Moreover, long-term electricity price forecasting can be explored. The proposed methodology can also be applied in other time series forecasting data.

Acknowledgment

This work is funded by Universiti Malaya research grant from Malaysia under the project name 'Intelligent Price Forecasting System for Optimal Energy Market' with grant number ST005-2021.

References

[1] A. Pourdaryaei, H. Mokhlis, H. A. Illias, S. H. A. Kaboli, and S. Ahmad, "Short-term electricity price forecasting via hybrid backtracking search algorithm and ANFIS approach," *IEEE Access,* vol. 7, pp. 77674-77691, 2019.

[2] I. Ozer, S. B. Efe, and H. Ozbay, "A combined deep learning application for short term load forecasting," *Alexandria Engineering Journal,* vol. 60, pp. 3807-3818, 2021.

[3] E. Almeshaiei and H. Soltan, "A methodology for electric power load forecasting," *Alexandria Engineering Journal,* vol. 50, pp. 137-144, 2011.

[4] A. Pourdaryaei, H. Mokhlis, H. A. Illias, S. H. A. Kaboli, S. Ahmad, and S. P. Ang, "Hybrid ANN and artificial cooperative search algorithm to forecast short-term electricity price in de-regulated electricity market," *IEEE Access,* vol. 7, pp. 125369-125386, 2019.

[5] G.-F. Fan, X. Wei, Y.-T. Li, and W.-C. Hong, "Forecasting electricity consumption using a novel hybrid model," *Sustainable Cities and Society,* vol. 61, p. 102320, 2020.

[6] Y.-Y. Hong, J. V. Taylar, and A. C. Fajardo, "Locational Marginal Price Forecasting Using Deep Learning Network Optimized by Mapping-Based Genetic Algorithm," *IEEE Access,* vol. 8, pp. 91975-91988, 2020.

[7] F. Wu, C. Cattani, W. Song, and E. Zio, "Fractional ARIMA with an improved cuckoo search optimization for the efficient Short-term power load forecasting," *Alexandria Engineering Journal,* vol. 59, pp. 3111-3118, 2020.

[8] H. Wang, Y. Liu, B. Zhou, C. Li, G. Cao, N. Voropai, *et al.*, "Taxonomy research of artificial intelligence for deterministic solar power forecasting," *Energy Conversion and Management,* vol. 214, p. 112909, 2020.

[9] Y. Shuai, T. Song, and J. Wang, "Integrated parallel forecasting model based on modified fuzzy time series and SVM," *Journal of Systems Engineering and Electronics,* vol. 28, pp. 766-775, 2017.

[10] J. H. Zhao, Z. Y. Dong, X. Li, and K. P. Wong, "A framework for electricity price spike analysis with advanced data mining methods," *IEEE Transactions on Power Systems,* vol. 22, pp. 376-385, 2007.

[11] J. Dhillon, S. A. Rahman, S. U. Ahmad, and M. J. Hossain, "Peak electricity load forecasting using online support vector regression," in *2016 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, 2016, pp. 1-4.

[12] X. Yan, Y. Song, and N. A. Chowdhury, "Performance evaluation of single SVM and LSSVM based forecasting models using price zones analysis," in *2016 IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC)*, 2016, pp. 79-83.

[13] A. Heydari, F. Keynia, D. A. Garcia, and L. De Santoli, "Mid-term load power forecasting considering environment emission using a hybrid intelligent approach," in *2018 5th International Symposium on Environment-Friendly Energies and Applications (EFEA)*, 2018, pp. 1-5.

[14] R. Zhang, G. Li, and Z. Ma, "A Deep Learning Based Hybrid Framework for Day-Ahead Electricity Price Forecasting," *IEEE Access,* vol. 8, pp. 143423-143436, 2020.

[15] S. Zhou, L. Zhou, M. Mao, H.-M. Tai, and Y. Wan, "An optimized heterogeneous structure LSTM network for electricity price forecasting," *IEEE Access,* vol. 7, pp. 108161-108173, 2019.

[16] P. Lv, S. Liu, W. Yu, S. Zheng, and J. Lv, "EGA-STLF: A hybrid short-term load forecasting model," *IEEE Access,* vol. 8, pp. 31742-31752, 2020.

[17] H. Manner, F. A. Fard, A. Pourkhanali, and L. Tafakori, "Forecasting the joint distribution of Australian electricity prices using dynamic vine copulae," *Energy Economics,* vol. 78, pp. 143-164, 2019.

[18] Y. Elfahham, "Estimation and prediction of construction cost index using neural networks, time series, and regression," *Alexandria Engineering Journal,* vol. 58, pp. 499-506, 2019.

[19] G. Hamilton, A. Abeygunawardana, D. P. Jovanović, and G. F. Ledwich, "Hybrid model for very short-term electricity price forecasting," in *2018 IEEE Power & Energy Society General Meeting (PESGM)*, 2018, pp. 1-5.

[20] M. Alazab, S. Khan, S. S. R. Krishnan, Q.-V. Pham, M. P. K. Reddy, and T. R. Gadekallu, "A multidirectional LSTM model for predicting the stability of a smart grid," *IEEE Access,* vol. 8, pp. 85454-85463, 2020.

[21] S. K. Gupta, M. Tripathi, and J. Grover, "Hybrid optimization and deep learning based intrusion detection system," *Computers and Electrical Engineering,* vol. 100, p. 107876, 2022.

[22] S. K. Gupta, M. Tripathi, and J. Grover, "Towards an Effective Intrusion Detection System using Machine Learning techniques: Comprehensive Analysis and Review," in *2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, 2021, pp. 1-6.

[23] S. Patidar, M. Tripathi, and S. K. Gupta, "Leveraging LSTM-RNN combined with SVM for Network Intrusion Detection," in *Proceedings of the International Conference on Data Science, Machine Learning and Artificial Intelligence*, 2021, pp. 26-31.

[24] M. R. Haq and Z. Ni, "A new hybrid model for short-term electricity load forecasting," *IEEE Access,* vol. 7, pp. 125413-125423, 2019.

[25] J. Bedi and D. Toshniwal, "Deep learning framework to forecast electricity demand," *Applied energy,* vol. 238, pp. 1312-1326, 2019.

[26] U. Ugurlu, I. Oksuz, and O. Tas, "Electricity price forecasting using recurrent neural networks," *Energies,* vol. 11, p. 1255, 2018.

[27] U. Ugurlu, O. Tas, A. Kaya, and I. Oksuz, "The financial effect of the electricity price forecasts' inaccuracy on a hydro-based generation company," *Energies,* vol. 11, p. 2093, 2018.

[28] C. Fan, Y. Sun, Y. Zhao, M. Song, and J. Wang, "Deep learning-based feature engineering methods for improved building energy prediction," *Applied energy,* vol. 240, pp. 35-45, 2019.

[29] A. S. Weigend, *Time series prediction: forecasting the future and understanding the past*: Routledge, 2018.

[30] P. J. Brockwell, P. J. Brockwell, R. A. Davis, and R. A. Davis, *Introduction to time series and forecasting*: Springer, 2016.

[31] J. Osborne, "Improving your data transformations: Applying the Box-Cox transformation," *Practical Assessment, Research, and Evaluation,* vol. 15, p. 12, 2010.

[32] A. Pal and P. Prakash, *Practical time series analysis: master time series data processing, visualization, and modeling using python*: Packt Publishing Ltd, 2017.

[33] I. E. Livieris, S. Stavroyiannis, E. Pintelas, and P. Pintelas, "A novel validation framework to enhance deep learning models in time-series forecasting," *Neural Computing and Applications,* vol. 32, pp. 17149-17167, 2020.

[34] A. Naz, N. Javaid, M. Asif, M. U. Javed, A. Ahmed, S. M. Gulfam, *et al.*, "Electricity Consumption Forecasting Using Gated-FCN With Ensemble Strategy," *IEEE Access,* vol. 9, pp. 131365-131381, 2021.

[35] T.-Y. Kim and S.-B. Cho, "Predicting residential energy consumption using CNN-LSTM neural networks," *Energy,* vol. 182, pp. 72-81, 2019.

[36] X. Xie, M. Li, and D. Zhang, "A Multiscale Electricity Price Forecasting Model Based on Tensor Fusion and Deep Learning," *Energies,* vol. 14, p. 7333, 2021.

Rashed Iqbal

Rashed Iqbal received his B.Sc. degree in electrical and electronic engineering from the BRAC University, Bangladesh, in 2016, currently he is pursuing his degree in Master of Engineering Science in power system from the Universiti Malaya (UM), Malaysia. His current research interests include artificial intelligence in power system, renewable energy, smart and microgrids.

Hazlie Mokhlis

Prof. Ir. Dr Hazlie Mokhlis is currently a Professor at Department of Electrical Engineering, Universiti Malaya. His research interests include power system analysis (distribution automation), power quality, and fault location. His detail CV can be obtained from https://umexpert.um.edu.my/hazli

Anis Salwa Mohd Khairuddin

Anis Salwa Mohd Khairuddin is currently working as a senior lecturer at Department of Electrical Engineering, Faculty of Engineering, Universiti Malaya, Malaysia. Her research interests include Expert system (machine learning, optimization, prediction), signal and image processing. Her detail CV can be obtained from https://umexpert.um.edu.my/anissalwa.

Munir Azam Muhammad

Munir Azam Muhammad received a B.E. degree in electronic engineering from Hamdard University in 2010 and an M.E. degree in Industrial Electronics from N.E.D University, Pakistan, in 2013. He is currently pursuing a Ph.D. in Electrical Engineering in the University of Malaya (UM), Malaysia. His research interests include distribution automation, load shedding, distributed generation and development of optimization strategies for a smart grid.

Design of a Machine Learning-Based Platform for Currency Market Prediction: A Fundamental Design Model

K. Gordillo-Orjuela, P. A. Gaona-García*, C. E. Montenegro-Marín*

Faculty of Engineering, Universidad Distrital Francisco José de Caldas, Bogotá (Colombia)

* Corresponding author: cemontenegrom@udistrital.edu.co (C. E. Montenegro-Marín), pagaonag@udistrital.edu.co (P. A. Gaona-García)

Received 22 February 2024 | Accepted 22 October 2024 | Published 25 November 2024

Abstract

Prediction models in foreign exchange markets have been very popular in recent years, and in particular, through the use of techniques based on Machine Learning. This growth has made it possible to train several techniques that increasingly allow us to improve predictions according to the criteria that each algorithm supports and can cover. However, the development of these models and their deployment within computer platforms is a complex task, given the variety of approaches that each researcher uses based on the training process and therefore by definition of the model, which leads to the consumption of high computing resources for its training, as well as various processes for its deployment. For this reason, the following article focuses on designing a technological platform oriented to micro services, which minimizes the consumption of resources and facilitates the integration of various techniques and the analysis of various criteria, which improves their analysis and validation in a Web environment.

Keywords

Forex, Machine Learning, Micro Services, Reduction Model, Software Architecture, Support Vector Machine (SVM).

DOI: 10.9781/ijimai.2024.11.002

I. Introduction

In recent years, artificial intelligence and in particular the area of Machine Learning (ML) have given rise to a large amount of research on price prediction in financial markets, such as hydrocarbons, precious metals, currencies, among others [1]. Currently, a series of models and algorithms have been proposed that seek to predict the value of different currencies in the foreign exchange market, also known as "Forex".

With the growing importance of artificial intelligence, prediction models based on machine learning have been developed, within which we can find seven broad categories: regression methods, optimization techniques, support vector machines (SVMs), neural networks, chaos theory, pattern-based methods, and other methods that include natural language processing [2].

Despite the development and rise of studies based on prediction techniques and models to analyze behaviors and characteristics of algorithms based on machine learning, the process of choosing, analyzing and implementing them within an application scenario is a complex process given the variety of resources, criteria and computational aspects that are required to simulate them within a work environment. Therefore, the design of a computer platform that allows the comparison of machine learning techniques could be of great value and usefulness to carry out the choice of models that allow the optimization and prediction of the value of currencies within specific scenarios and, in turn, allow the combination of some of these models to obtain even more accurate results.

The objective of the following article is to propose the design of a software platform that allows the execution of several models based on Machine Learning, to carry out prediction processes in the prices of the foreign exchange market, in order to identify elements that facilitate their implementation within various scenarios and improve the process of deployment and measurement of these models in various scenarios.

The rest of the article is divided as follows: in Section II the background is addressed where the different methods, techniques and mechanisms that have been developed for prediction within the foreign exchange market will be analyzed, in order to identify which of them can be implemented within a computational platform. Section III presents the methodology that was used to carry out the design of the computational platform, Section IV proposes the design of the computational platform based on a model based on software architecture. Section V presents the architectural approach. Section VI presents the results of this platform design and some results that were obtained from its development. Finally, section VII presents the findings and future work.

Please cite this article as:
K. Gordillo-Orjuela, P. A. Gaona-García, C. E. Montenegro-Marín. Design of a Machine Learning-Based Platform for Currency Market Prediction: A Fundamental Design Model, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 1, pp. 162-172, 2024, http://dx.doi.org/10.9781/ijimai.2024.11.002

II. Background

A. Forex Currency Study

The foreign exchange market, also known as "FOREX", is a transactional mechanism where a part of the population interested in this mechanism can acquire units of one currency to buy a proportional amount in another. There are two types of analysis on this market: fundamental analysis and technical analysis. In this paper, we will focus on technical analysis, which, in most cases, is based on statistical graphs and machine learning techniques [3].

Other proposals for price prediction in the foreign exchange market focus on improving existing machine learning algorithms through the use of optimization techniques [4]. Some research also uses genetic algorithms, such as particle swarm optimization (PSO). For example, in the study conducted by Pradeep kumar [5], the PSO algorithm was used to train a quantile regression neural network, allowing predicting the value of the forex market for USD/INR, EUR/USD, and USD/JPY. The research yielded good results based on the EM and MSE, although only for certain currencies and not for all currencies used in the study in general.

B. Analysis Tools and Algorithms Based on Machine Learning

Although numerous studies have been conducted on this topic, each with promising results, little research has been done on the development of a platform that allows several of these models to be collected and run in an environment that allows for statistical comparison. Within the literature we can find several proposals where the authors focus on the exchange and execution of models based on Machine Learning in a general way, but not in a specific context such as within the FOREX market. Fig. 1 describes 4 related works that are close to the approach that is being carried out through proposals based on Machine Learning.

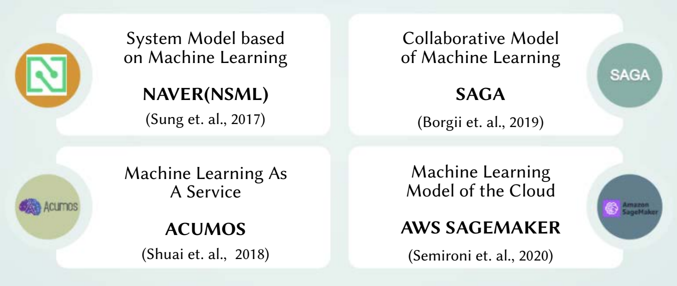


Fig. 1. Related Jobs Based on Machine Learning.

The NAVER Smart Machine Learning (NSML) platform, introduced by [6], aims to help researchers focus on modeling ML systems, rather than performing low-level tasks such as monitoring the training state of neural networks or mapping graphics processing units. The system consists of three main parts: a task scheduler, which efficiently allocates memory to execute ML tasks; a container system that provides a runtime and storage environment for models running on the system; and a graphical interface that acts as a mediator between the outputs generated by the system and the input flows provided by users.

Although there are models and studies on forex market prediction, there are few initiatives that focus on collecting, running, and comparing such models on a software platform. One such initiative is SAGA, an open-source platform presented by Borgli [7]. SAGA allows you to share ML models, training techniques for neural networks, and datasets. In addition, the platform offers the possibility of mixing several models and including extensions to improve the training of neural networks. This initiative promotes collaboration and knowledge sharing in the field of forex market forecasting.

According to Zhao [8], machine learning models are often developed for specific tasks using different frameworks, which makes it difficult to reuse and improve them. To address this issue, they introduce the open-source platform called ACUMOS, which allows ML models to be packaged into micro services containers and shared through the platform's catalog. Although this platform facilitates the reuse of models by treating them as "black boxes" and focusing on communication, it does not allow model optimization and leaves this task to the developers.

Some models focus on improving or combining ML machine learning techniques with natural language processing. This is the case of the work carried out by Semiromi [9] where the authors combine the use of textual information from news from the economic calendar with indicators and historical information from the foreign exchange market to train ML models such as extreme gradient boosting (XGB), random forest (RF) and SVM. At the end of the study, statistical measurement accuracy was above 60%, and in some cases 64%, for predictions in the following thirty minutes. According to the authors, this shows that text mining combined with ML techniques can improve prediction accuracy in FOREX.

Based on this panorama and in order to compare and allow the selection of a model based on Machine Learning that best adapts to the prediction of the currency selected for our case study, a platform is proposed that facilitates the execution of several of these prediction algorithms and provides certain useful metrics that provide relevant information for decision-making regarding the purchase of a certain currency.

In this sense, there are authors who propose an approach to build an ensemble classifier using sentiment in news at sentence level and technical indicators to predict stock trends [10], but because it is unstructured data, extracting the most important features is difficult. Moreover, positive or negative news does not always affect stock prices in a certain way, so it is a purpose not viable to this work.

Finally, the time series approach could be used for forecasting, as done is some works [11]: autoregressive integrated moving average (ARIMA) and seasonal autoregressive integrated moving average (SARIMA) to forecast the impact of COVID-19 on cigarette sales. This could be the basis to apply it to currency market prediction, for now we will focus only on Machine Learning-based techniques.

III. Research Methods

This research aims to propose the software design of a platform that allows the integration of multiple models of price prediction of the foreign exchange market. Fig. 2 shows the phases that were addressed for the proposal.

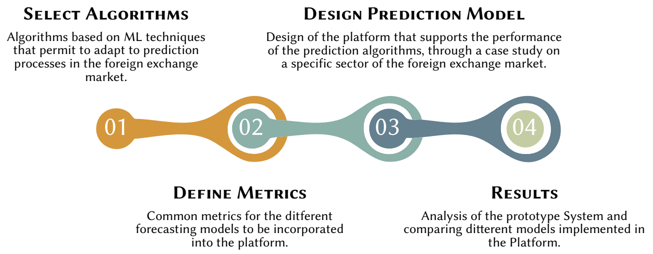

Select Algorithms
Algorithms based on ML techniques that permit to adapt to prediction processes in the foreign exchange market.
Design Prediction Model
Design of the platform that supports the performance of the prediction algorithms, through a case study on a specific sector of the foreign exchange market.
01
02
03
04
Define Metrics
Common metrics for the ditferent forecasting models to be incorporated into the platform.
Results
Analysis of the prototype System and comparing ditterent models implemented in the Platform.

Fig. 2. Research methods.

Fig. 2 shows 4 phases of work. Phase 1, oriented to the selection of algorithms, will be based on a literature review addressed in the background section of the most common uses algorithms with the best results for price prediction in the foreign exchange market. This will make it possible to establish the inputs for the design of the application, the input and output values of the models, as well as to identify the statistical measures necessary to compare the performance of these algorithms.

In the next phase (definition of metrics), metrics will be identified to compare the performance of different ML models for prediction in the forex market. This involves defining the output interfaces of each algorithm and the mechanisms for storing and calculating errors in the predictions, in line with the literature identified in the background section. Since each model behaves differently, it will be necessary to perform a preliminary design of the communication mechanism between the application and the models.

The third phase of the project will focus on the design of the software platform that will allow the execution and comparison of the selected algorithms at a statistical level. In this phase, the definition of the architecture to be used, the selection of the necessary software components, the definition of the communication interfaces, the design of the graphical user interface and the creation of the mechanism to support the execution of the algorithms, which can be developed using multiple frameworks, will be carried out.

For the last phase, the implementation of a prototype of the designed solution will be carried out and then the results of selected algorithms will be presented. For software development, the benefits and adaptations of the Kanban agile development methodology will be used [12] due to its characteristics of flexibility and adaptability to changes.

IV. Platform Design

A. Analysis and Selection of Algorithms

For practical purposes of prototype design, the use of five Machine Learning techniques was determined, following the approach adopted in other studies such as those of Bansal [13] and Biswas [14]. Below, the mode of operation to carry out its implementation within the prototype of the platform is briefly described.

1. SVR-Wavelet Adaptive Model for Forecasting Financial Time Series

This work was carried out by Raimundo [15]. The study focuses on the use of regression support vectors (SVRs) and the Wavelet transform to make predictions in financial series over time, specifically in the foreign exchange market (FOREX). SVRs are a mathematical model used in Machine Learning that allows regression analysis and data classification. This model, according to the authors, has proven to perform better than the ARIMA and ARFIMA models in terms of predicting the price of the Australian dollar against the Japanese yen. To compare the models, several metrics were used, including ME (Mean Error), RMSE (Root Mean Square Error), MAE (Mean Absolute Error), MPE (Mean Percent Error), and ASM (Mean Absolute Percentage Error).

2. Bayesian Compressed Vector Auto Regression for Financial Time-series Analysis and Forecasting

This work published by Paponpat [16] has its origins in the previous work published by G. Koop, D. Korobilis and D. Petternuzzo [17], in which the autoregressive vector model (VAR) is combined with the Bayesian compression model (BC). VAR models are commonly used for prediction in financial markets. However, when the number of variables increases, the computational load also increases. This problem, known as the "dimensionality problem," occurs when the number of predictors in the VAR model equations is greater than the number of observations. In other words, "high-dimensionality methods" are used to refer to models that deal with this problem.

3. CNN–LSTM Model for Gold Price Time-series Forecasting

Proposal put forward by Livieris [18] where neural networks are used to predict the price of gold in the stock market. Specifically, two machine learning models are combined: convolutional neural networks (CNNs) and short-long term memory (LSTM) recurrent neural networks. CNNs are often used in image processing, as their layered architecture allows image data to be broken down into smaller components and analyzed using filters. Unlike CNN networks, recurring networks allow you to process data streams such as videos, written text, voice, time series, among others, as this data makes sense when analyzed together. For example, when analyzing a text, it is not enough to interpret word by word or letter by letter, but it is necessary to analyze how these words/letters are concatenated to make sense of a sentence or phrase.

To analyze the correlation between the data, recurrent neural networks not only use the activation of the current neuron, but also use the activation of the previous iteration. In other words, such networks implement a kind of "memory" that allows them to analyze the results of a previous activation.

As shown in Fig. 3, the neurons in an LSTM network are composed of three gates that control the flow of information to and from the "C" state cell, which can be understood as a band that stores the relevant information of the network. These gates act as valves that allow the state cell to be modified, either to forget, add, or give an output. The first gate is the "Forget gate", which allows you to remove unnecessary information from the status cell. The second is the Input gate, which allows you to add new information to the state cell. Finally, the Output gate allows us to generate the output of the LSTM neuron.

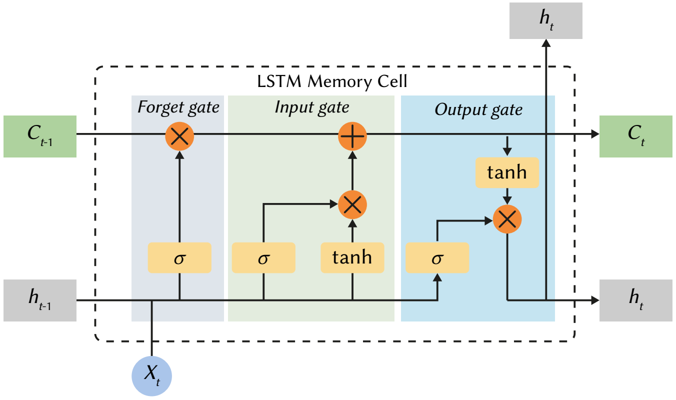


Fig. 3. Unit of an LSTM neuron (Source: [19] under license CC BY https://creativecommons.org/licenses/by/4.0/) .

4. Stock Price Forecasting Via Sentiment Analysis on Twitter

In this work [20], a large number of tweets from different dates related to the price of a certain stock market stock are collected. Then, using sentiment analysis techniques and the use of SVM (Support Vector Machine), a prediction is made of the price of the stock being analyzed.

The module works relatively simply: first, a message cleanup is performed to remove special characters and non-relevant information, such as emojis. Then, each message is scored based on the emotion it expresses, using keywords. For example, if a message contains words such as "bad," "hate," or "disgust," it is considered to have a negative emotionality score. On the other hand, if a message uses words like "love," "opportunity," "excellent," or "good," it is considered to have a positive emotionality. Connectors, such as "to", "from", "the", among others, are omitted. In the end, each message is weighted and labeled as positive, negative, or neutral. This information is put into a matrix, along with the date of the tweet and moves on to the next phase, which is processing and analysis.

When you have the sentiment score matrix and the time series matrix, you proceed to merge both into a single characteristic matrix. This matrix of characteristics is constructed with the following parameters for each of the days prior to the prediction: percentage of positive sentiments, percentage of negative sentiments, percentage of neutral sentiments, closing price of the stock/currency, HLPCT (High-Low percentage), PCTchange (Percentage Change), volume of the stock/currency.

5. Price Forecasting for Agricultural Products Based on BP and RBF Neural Network

In this work carried out by Yu, Shouhua and Ou, Jingying [21], neural networks based on radial functions (RBF) and the backpropagation technique (Backpropagation) are used to predict the value of certain agricultural products in the Chinese market, which was tested between January and December 2011.

B. Criteria for Analysis

Within the literature review carried out by Islam (2020) [2], seven main categories can be identified in which the current advances in machine learning applied to prediction in the foreign exchange market can be classified. What all of these works have in common is the use of statistical measures to establish the accuracy of their predictions. Although the measures used may vary depending on the model approach, the following statistical measures are generally employed: MAE (Mean Absolute Error), MSE (Mean Square Error), RMSE (Root-meansquare Error), and ASM (Mean Absolute Percentage Error). All these measures seek to show how far the prediction is from the actual value in the market.

Based on the algorithm review presented in the previous section and considering the conceptual framework, three main metrics have been selected for the comparison of the five selected algorithms: MAE (Mean Absolute Error), RMSE (Root Mean Square Error) and ASM (Mean Absolute Percentage Error).

It was decided to use these three measures because, according to the literature review described in the background section, they are the most widely used to present the results and allow the degree of error of each algorithm in the prediction to be determined. In addition, authors such as Alexei Botchkarev [22] highlight that these metrics have been the most popular in the last 25 years. The mathematical formulation of each of these measures is presented below.

1. MAE

MAE seeks to give a measure of how far the calculated data is from the observed data. In other words, how far the prediction would be from reality. Its formula is given by (1):

$$MAE = \sum_{i=1}^{n} \frac{|y_i - x_i|}{n} \tag{1}$$

Where: y_i, is the value of the prediction and x_i is the actual or observed value.

2. RSME

The Root Mean Square Error (RMSE), like the MAE, allows you to measure how far the predicted values are from the observed or actual values. The main difference between these two metrics is that the RMSE shows us the squared difference between the estimated value and the observed value.

In simple words, the RMSE tells us how accurate a model is at representing the phenomenon of reality. Its mathematical formulation tells us how far the data being predicted is off the ground from the actual data expressed in (2)

$$\text{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(Q_i - y)^2}{n}} \tag{2}$$

Where n is the number of observations. $\hat{y}_i$, is the predicted value at instant i. y_i, is the real value at instant i.

The MSE alone penalizes errors that are larger in the model. On the other hand, the RMSE, by taking the square root of the MSE, allows us to have a better perspective in terms of scoring when there are some very large errors that, at a general level, should not impact the performance of the model.

3. ASM

The Mean Absolute Percentage Error (ASM) is a measure that allows us to express the absolute error in percentage terms. It gives us an estimate of the accuracy of the prediction method we are using. It is more intuitive to understand than absolute error since error is expressed in percentage terms for a given time series.

For example, if we have an ASM value of 5%, it is interpreted to mean that the difference between the predicted value and the actual value is approximately 5%. Its formulation is simple and is expressed in (3).

$$MAPE = \frac{100\%}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right| \tag{3}$$

Where n is the number of measurements, A_t it's a real number. F_t, is the value predicted by the algorithm.

Based on these criteria, the following section focuses on a description of the architectural approach made.

V. Architectural Approach

A. Conceptual Model

From the development perspective, the technical team will be in charge of adding new algorithms to the platform, conceptually following the process described in Fig. 4, where the possibility of the platform taking the code from a repository and performing the provisioning and cataloging of said module within the system is raised.

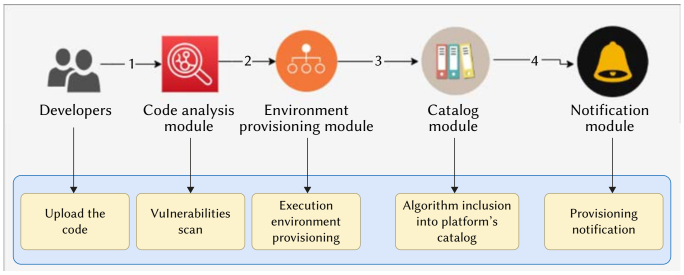

Developers
1
Code analysis module
2
Environment provisioning module
3
Catalog module
4
Notification module
Upload the code
Vulnerabilities scan
Execution environment provisioning
Algorithm inclusion into platform's catalog
Provisioning notification

Fig. 4. Conceptual model of the process that a developer follows to link a new algorithm, prediction with ML, to the platform.

The steps to link a new algorithm are: (1) The developer uploads the code to the repo and starts the algorithm linking process. (2). Once the vulnerability scan is completed, the environment provisioning module is triggered. (3). After the provisioning module is done, there are an initial call to the algorithm to check is functionality. (4). Once the algorithm is added to the catalog, there is a notification sent to the developer.

However, from the perspective of a user who needs to analyze the performance of the different algorithms in the catalog, the platform should allow the selection of algorithms from the catalog to which the user has access and run them with different data sets to show the results in relation to three metrics: MAE, RMSE and ASM. This is how the person who requires it, makes a prediction of "n" days in the future for the time series that is entered as a parameter. This behavior is described in Fig. 5.

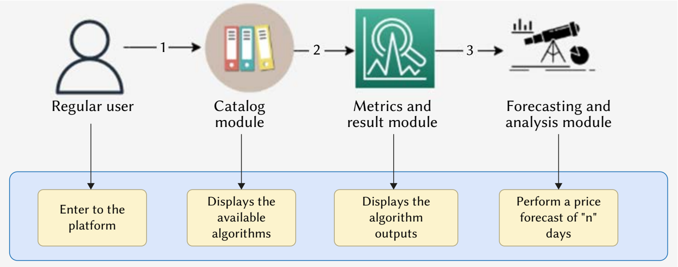

Regular user
1
Catalog module
2
Metrics and result module
3
Forecasting and analysis module
Enter to the platform
Displays the available algorithms
Displays the algorithm outputs
Perform a price forecast of "n" days

Fig. 5. Conceptual model of the process that a regular user follows to use and analyze prediction models with ML within the platform.

The steps for the conceptual model of the process that a regular user follows are: (1). The regular user selects one or more algorithms from catalog to compare. (2). The metrics module displays the results of each algorithm for the analysis. (3). Once the results are displayed, the user can select any of the algorithms and ask for a price forecast.

B. Process Model

For the incorporation of a new algorithm to the platform's catalog, the business process shown in Fig. 6 is followed.

The diagram presents the request to add a new algorithm to the platform via a link to the code repository. The system then performs a scan for potential vulnerabilities in the code. If any vulnerabilities are detected, the user is notified. Otherwise, the runtime environment is provisioned and an initial execution test is performed. Finally, if the execution was successful, the algorithm is added to the catalog and the requestor is notified.

C. Proposed Architecture

This section focuses on defining the application architecture and how it works at a high level. Both, the architecture follows the architectural pattern of publisher/subscriber and the use of machine learning models we include in this repository https://github.com/kgordillo-hub/SVM-Wavelet-forecasting-Financial-Time-Series.

Table I presents a description of the code developed for the assembly of the prototype. The web services were written in Java and the user interface in ReactJS.

TABLE I. Repository Used to Develop and Assembly the Platform

Name	Link
Code analyzer	https://github.com/kgordillo-hub/AnalizadorCodigo
Environment generator	https://github.com/kgordillo-hub/GeneradorEntornos
API Parameters manager	https://github.com/kgordillo-hub/GestorParametrosAPI
Metrics	https://github.com/kgordillo-hub/Metricas
Catalog manager	https://github.com/kgordillo-hub/GestorCatalogo
Results manager	https://github.com/kgordillo-hub/GestionResultados
Code linker	https://github.com/kgordillo-hub/VinculadorCodigo
Frontend	https://github.com/kgordillo-hub/material-kit-react
Wavelets+SVR	https://github.com/kgordillo-hub/1.Wavelets_SVM
BCVAR	https://github.com/kgordillo-hub/2.BCVAR
LSTM+CNN	https://github.com/kgordillo-hub/3.LSTM_CNN
SA + SVR	https://github.com/kgordillo-hub/4.SA_Twitter
RBF + NN	https://github.com/kgordillo-hub/5.RBFNN

By using the event bus, each service subscribes and receives messages as needed. In this way, the asynchronous operation of the services is guaranteed, and they are decoupled, this deployment is presented in Fig. 7.

As can be seen in Fig. 7, from left to right, the user accesses the platform from a terminal and communicates with the identity provider to obtain a JWT token, entering a username and password, following the OpenID Connect (OIDC) protocol and OAuth2.0. Once the client obtains the token, it includes it in the header of the HTTPS request to access the application. All incoming requests, both to the web layer and the composition services layer, will be made through a gateway. Each component that was defined for its construction is described below.

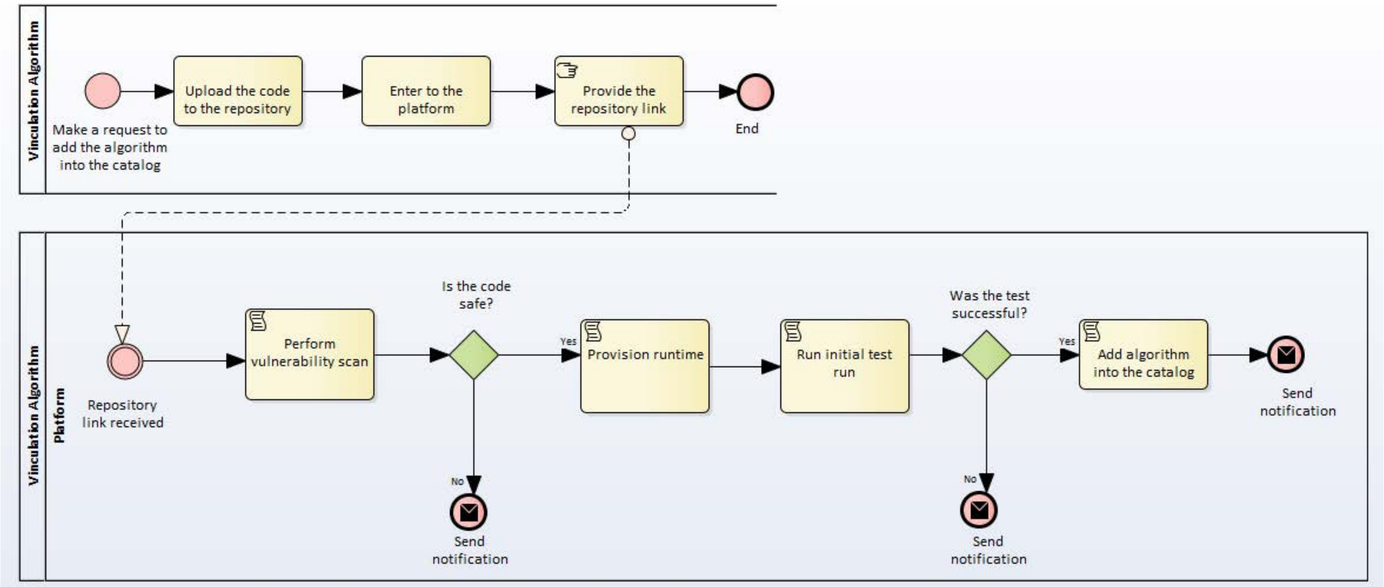


Fig. 6. Business process model for linking a new ML algorithm to the platform.

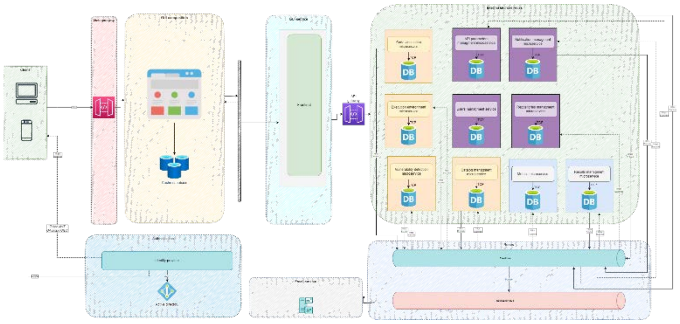


Fig. 7. Architecture proposed for the prototype using event bus and micro services.

1. Security and Authentication

Since the code is going to be included in the platform, a mechanism for scanning for possible vulnerabilities must be established. It is necessary to have a service in charge of performing a static analysis of the code in search of possible security breaches and storing only the code that does not present vulnerabilities. For this, there are multiple providers that offer security analysis tools, each with its own integration parameters and format of the rules that are loaded. Some of the functionalities that the administrator will have access to are adding, editing and deleting code analysis providers. Additionally, you can manage the code analysis rules that are loaded for each provider. There may be cases in which code analysis is not necessary or not possible. For such cases, there is the possibility of disabling/enabling mandatory code scanning by user groups. For these aspects, a microservice called Code Vulnerability Detection Microservice was developed, which can be identified in orange in Fig. 7.

Related to the authentication layer there are mainly two components. The first component is the identity provider, which will be responsible for validating the user's identity and returning a JWT token using the OAuth2.0/OIDC authentication protocol. This token will have the permissions of the services that the user can access. This repository will store the data of the users, the information of the groups to which they belong and the permissions they have to consume the different services of the system. Fig. 7 presents this service in green color.

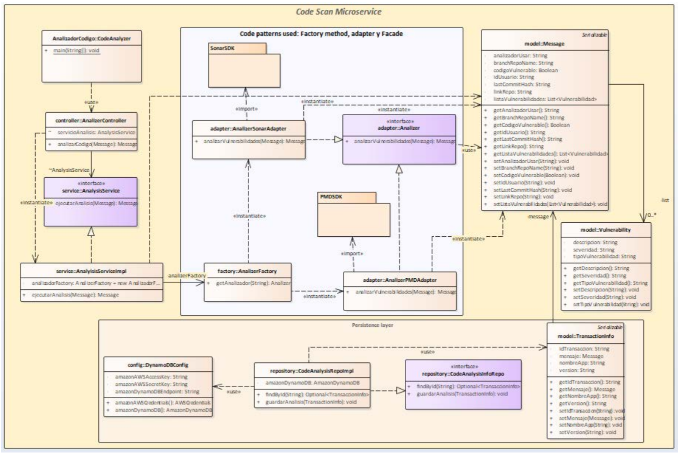
2. IGU Composition Service

The web part or "frontend" will have its own cached database to maintain the states that will be displayed to the user. On the other hand, in the "GUI (Graphical User Interface) compositing services" layer, there will be a single component that will allow the composition of the visual part. These components are described in Internal Micro services, which allows to support the logic of the application or "backend" will be composed of micro services that communicate through an event bus. Each micro service will have its own database to maintain the state and configuration of the service it provides. Below is a description of each of the functionalities of the services to be developed:

- **Code linking and execution micro services**: Within the architecture of Fig. 7, the services presented in orange are responsible for the management, configuration and execution of the algorithms to be linked on the platform, as well as the maintenance of the algorithms within the execution process.
- **Platform management micro services**: Likewise, the services presented in purple in Fig. 7 are responsible for the management of the platform's configuration, for example, the configuration of notifications or the catalog of algorithms, the parameters to be used for the consumption of each of them, successful linking and completion of algorithms used, among others.
- **Result micro services:** The services presented in light blue in Fig. 7 are responsible for the use of metrics and management of results produced by the different ML algorithms linked to the platform.

D. Structural Model

As shown in Fig. 8, the class diagram is sketched for the micro service in charge of analyzing the code to be linked for vulnerabilities. This service will be implemented using the "Spring Boot" framework. The structure will consist of a "Controller" class in charge of receiving and dispatching requests through the HTTP protocol, a class that implements the logic of the service that is offered, and a logical layer that contains the implementation of the code analysis.



Code Scan Microservice
Code patterns used: Factory method, adapter y Facade
AnalizadorCodigo::CodeAnalyzer
+ main(String[]): void
«use»
controller::AnalizerController
~ servicioAnalisis: AnalysisService
+ analizarCodigo(Message): Message
~AnalysisService
«interface»
service::AnalysisService
«instantiate»
ejecutarAnalisis(Message): Message
service::AnalyisisServiceImpl
analizadorFactory: AnalizerFactory = new AnalizadorF...
+ ejecutarAnalisis(Message): Message
analizerFactory
SonarSDK
«import»
«instantiate»
adapter::AnalizerSonarAdapter
+ analizarVulnerabilidades(Message): Message
«interface»
adapter::Analizer
+ analizarVulnerabilidades(Message): Message
«use»
PMDSDK
«import»
factory::AnalizerFactory
+ getAnalizador(String): Analizer
adapter::AnalizerPMDAdapter
+ analizarVulnerabilidades(Message): Message
Serializable
model::Message
analizadorUsar: String
branchRepoName: String
codigoVulnerable: Boolean
idUsuario: String
lastCommitHash: String
linkRepo: String
listaVulnerabilidades: List<Vulnerabilidad>
+ getAnalizadorUsar(): String
+ getBranchRepoName(): String
+ getCodigoVulnerable(): Boolean
+ getIdUsuario(): String
+ getLastCommitHash(): String
+ getLinkRepo(): String
+ getListaVulnerabilidades(): List<Vulnerabilidad>
+ setAnalizadorUsar(String): void
+ setBranchRepoName(String): void
+ setCodigoVulnerable(Boolean): void
+ setIdUsuario(String): void
+ setLastCommitHash(String): void
+ setLinkRepo(String): void
+ setListaVulnerabilidades(List<Vulnerabilidad>): void
list
0..*
message
model::Vulnerability
descripcion: String
severidad: String
tipoVulnerabilidad: String
+ getDescripcion(): String
+ getSeveridad(): String
+ getTipoVulnerabilidad(): String
+ setDescripcion(String): void
+ setSeveridad(String): void
+ setTipoVulnerabilidad(String): void
Persistence layer
config::DynamoDBConfig
amazonAWSAccessKey: String
amazonAWSSecretKey: String
amazonDynamoDBEndpoint: String
+ amazonAWSCredentials(): AWSCredentials
+ amazonDynamoDB(): AmazonDynamoDB
repository::CodeAnalysisRepoImpl
amazonDynamoDB: AmazonDynamoDB
+ findById(String): Optional<TransaccionInfo>
+ guardarAnalisis(TransactionInfo): void
«interface»
repository::CodeAnalysisInfoRepo
+ findById(String): Optional<TransaccionInfo>
+ guardarAnalisis(TransactionInfo): void
Serializable
model::TransactionInfo
idTransaccion: String
mensaje: Message
nombreApp: String
version: String
+ getIdTransaccion(): String
+ getMensaje(): Message
+ getNombreApp(): String
+ getVersion(): String
+ setIdTransaccion(String): void
+ setMensaje(Message): void
+ setNombreApp(String): void
+ setVersion(String): void

Fig. 8. Class diagram for the vulnerability scanning microservice in code.

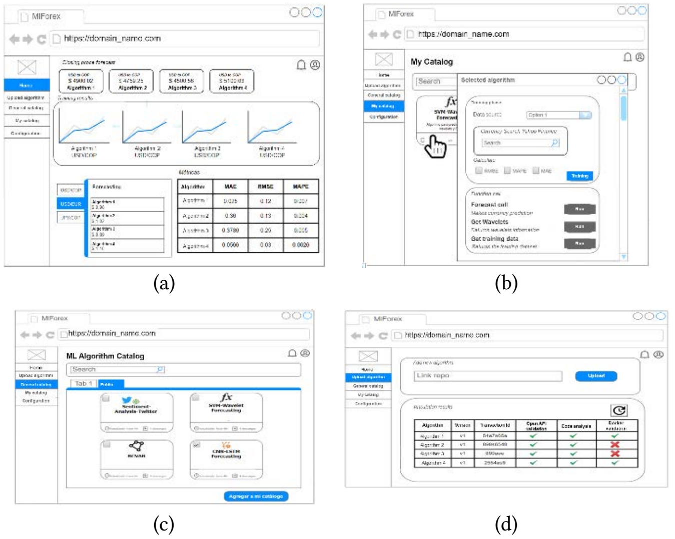
On the other hand, this service uses the façade pattern to hide the call to the different code analysis providers, such as PMD and Sonar. Because each vendor has different logic for each call to their implementations, the object adapter pattern is used to take the input message, transform it to the needs of each SDK (software development kit), and return a response to the service layer. Finally, the Factory Method pattern is implemented to obtain the different analysis adapters depending on the value that arrives in the Message object.

For the implementation of the services, Java was selected as the programming language and the "Spring framework" framework was selected to speed up the development process. On the other hand, for the implementation of the graphical user interface, the JavaScript programming language and the "ReactJS" framework were used. For the implementation of the databases and data bus, two technologies provided by Amazon Web Services (AWS) are used: DynamoDB and Simple Notification Service (SNS).

E. Front-end Graphic Prototype

Regarding graphic design, the following models were proposed to carry out its implementation, as shown in Fig. 9.

Fig. 9(a) shows the Web Schema of the page that allows you to display the general catalog of the system, in particular it allows to add algorithms to the personal catalog to analyze it. The interface in Fig. 9(b) allows you to interact with the algorithms added to the personal catalog to call the training method and the other methods offered by the selected ML algorithm. Fig. 9(c) presents the proposed Web schema that allows the main results of the algorithms to be displayed. Finally, Fig. 9(d) shows the status of the algorithms' linkage.



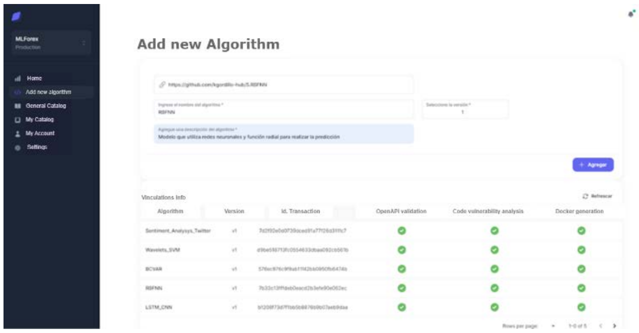
(a) (b) (c) (d)

Fig. 9. Architecture proposed for the prototype using event bus and micro services.

VI. Results

A. Architectural Software

The interface shown in Fig. 9 allows the user to enter the algorithm information to be linked, verify the link to the repository, model name, description, and version. Then, clicking "Add" allows you to perform model validation.



Add new Algorithm

https://github.com/kgordillo-hub/5.RBFNN

Ingrese el nombre del algoritmo *
RBFNN

Seleccione la versión *
1

Agregue una descripción del algoritmo *
Modelo que utiliza redes neuronales y función radial para realizar la predicción

Vinculations Info

Algorithm	Version	Id. Transaction	OpenAPI validation	Code vulnerability analysis	Docker generation
Sentiment_Analysys_Twitter	v1	7d2f92e0d0739dced91a77f26d31ffc7	✓	✓	✓
Wavelets_SVM	v1	d9be5f6713fc0554633dbae092cb567b	✓	✓	✓
BCVAR	v1	576ec976c9f9ab11f42bb0950fb6474b	✓	✓	✓
RBFNN	v1	7b33c13fffdeb0eacd2b3efe90e062ec	✓	✓	✓
LSTM_CNN	v1	b1208f73d7f1bb5b8876b9b07aeb9daa	✓	✓	✓

Fig. 10. Algorithm linking screen to the platform.

As presented in Fig. 10, the main flow of linking algorithms to the platform was implemented. To make use of this screen, the user must first specify the public URL of the repository where the algorithm to be run is hosted. Second, you need to add the name of the algorithm, its description, and the version to be linked, and click the "Add" button. After that, you must wait for the API specification structure validations, image generation, and vulnerability scanning to complete. Once the linking has been done, the algorithm will appear in the "General Catalog" section, as shown in Fig. 11. In this section, the user can click on "Add to my catalog" to add the algorithm to their list of available algorithms. This platform is designed for multiple users to link multiple algorithms and can be used collaboratively.

With the interface shown in Fig. 10, the user can view and select the models to use, which were described in the background section. Finally, Fig. 12 shows the results pane. This panel has two main sections: one that shows the estimated price for each algorithm for a specific currency, and another that shows the training result of each model and their respective comparison metrics.

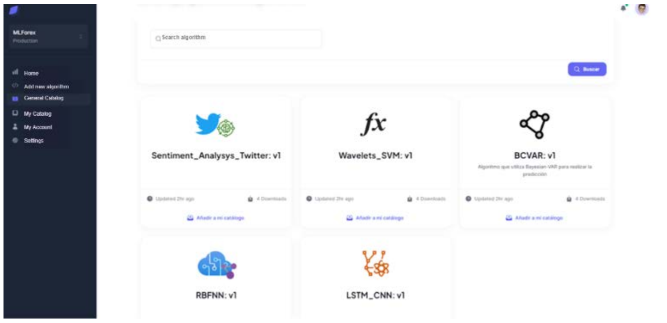


Fig. 11. General catalogue of algorithms linked to the platform.

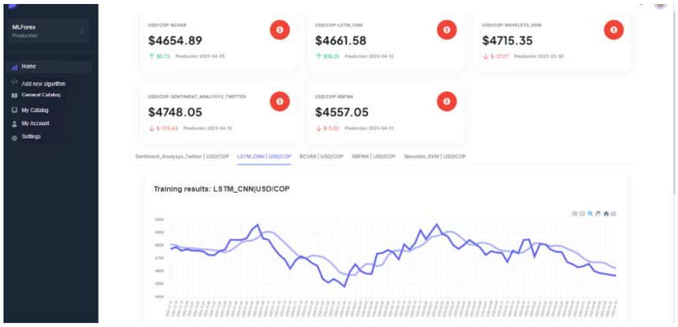


Fig. 12. Prediction results main screen.

B. Comparison of the Algorithms Selected

Following the first phase of the proposed methodology, the results of the five algorithms selected in the methodology section were replicated using the Python programming language; specifically, using Jupyter Notebooks as a tool. Once the algorithms were obtained, their performance was compared using the three metrics mentioned above (MAE, RMSE and MAPE), which, as Botchkarev [22] points out, allow us to understand how far the estimates are from the observed values, and have been the most used statistical measures since 1982. Below are the results of each algorithm for the following currency pairs: USD/EUR, USD/JPY, USD/COP and EUR/COP. The data evaluated correspond to the period from January 1, 2022, to August 1 of the same year (01/01/2022 - 01/08/2022).

TABLE II. Analysis of Results

Date	Open	High	Low	Close	Adj Close	Volume
2022-01-03	4557.200195	4571.700195	4557.200195	4557.200195	4557.200195	0
2022-01-04	4571.700195	4571.700195	4539.399902	4571.700195	4571.700195	0
2022-01-05	4539.399902	4568.399902	4539.399902	4539.399902	4539.399902	0
2022-01-06	4568.399902	4568.399902	4504.100098	4568.399902	4568.399902	0
2022-01-07	4504.100098	4510.799805	4504.100098	4504.100098	4504.100098	0
...	...	...	...	...	...	...
2022-07-26	4519.600098	4519.600098	4485.399902	4519.600098	4519.600098	0
2022-07-27	4485.399902	4485.399902	4462.200195	4485.399902	4485.399902	0
2022-07-28	4462.200195	4462.200195	4434.200195	4462.200195	4462.200195	0
2022-07-29	4434.200195	4434.200195	4426.399902	4434.200195	4434.200195	0
2022-08-01	4426.399902	4426.399902	4345.000000	4426.299902	4426.399902	0
151 rows x 6 columns						

The percentage of training data was 75%, leaving the remaining 25% as test data. That is, the time series from 01/01/2022 to 08/06/2022 was used as training information for each model, and the remaining time series, from 09/06/2022 to 01/08/2022, was used as test data to measure the accuracy of each algorithm, in accordance with the data sets presented at https://finance.yahoo.com/. An example of this data

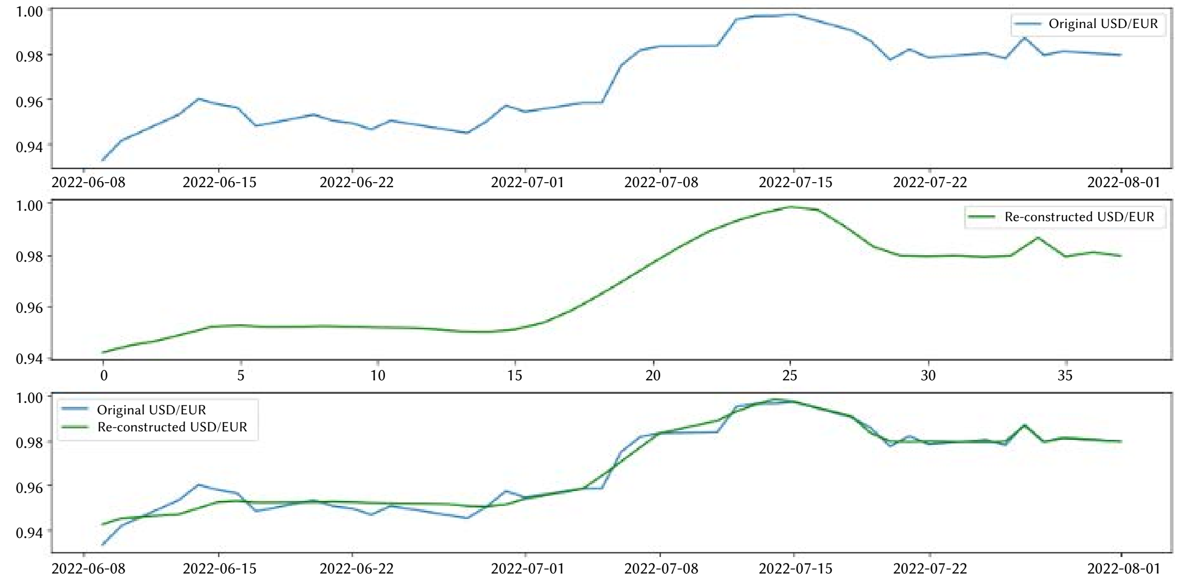


Fig. 13. Comparison graphic between actual and estimated values for USD/EUR using the SVR+Wavelet algorithm.

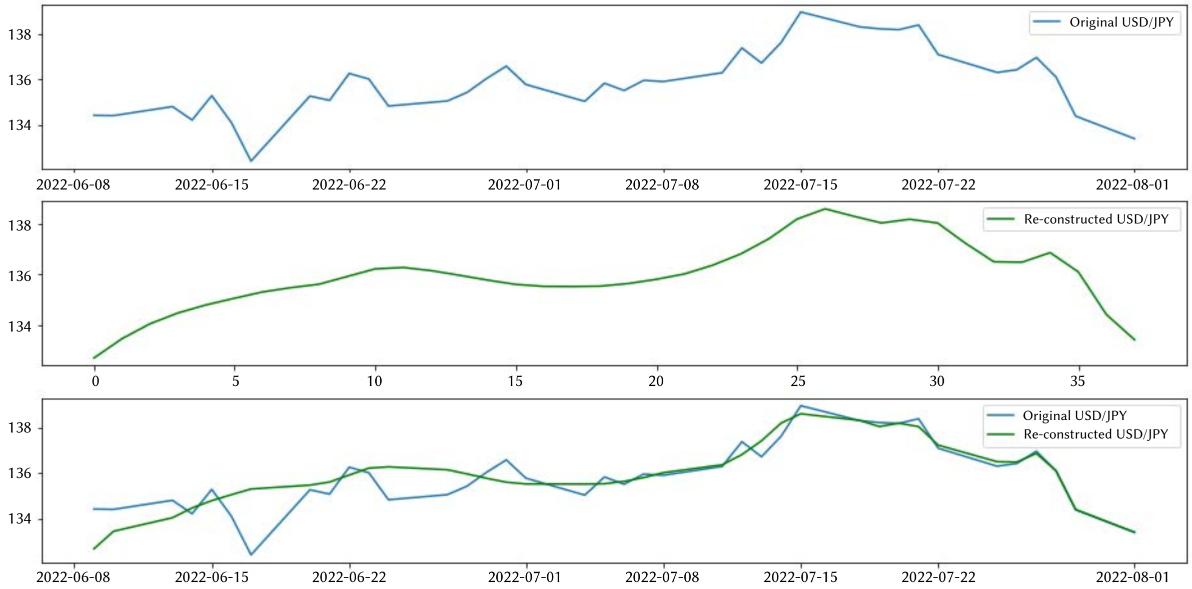


Fig. 14. Actual vs estimated value chart for USD/JPY using SVR+Wavelet algorithm. MAE: 0.485 RMSE: 0.746 MAPE: 0.35%.

is shown in Table II for the conversion of Euro (EUR) to Colombian peso (COP). It shows the opening prices, highest price, lowest price and closing price.

The following section present in detail the most promising results obtained of best algorithms selected, based on a case study defined for each currency pair during the period established above. At the end, a comparative table summarizing the performance of all algorithms is presented, in order to obtain an overview of all algorithms selected.

1. Performance: SVR-Wavelet Adaptive Model

The SVR-Wavelet algorithm, as described above, aims to decompose the time series into smaller components to remove noise and then take the cleanest component and perform prediction using SVR.

a) USD/EUR

The average value of the selected period of the training data, for the conversion of US dollars to euros, was 0.924 EUR for each USD. Performing the prediction, using the test data, the measurements shown in Fig. 13 are found.

The Y-axis shows the value of the currency on a scale of 0 to 1, while the X-axis shows the date on which the value was recorded. MAE: 0.0029 RMSE: 0.0039 MAPE: 0.30%

b) USD/JPY

The average value of the selected period for the training data, for the conversion of US dollars to Japanese Yen, was 125.00 YEN for every 1 USD. Performing the prediction, using the test data, the measurements shown in Fig. 14 are found.

As can be seen in Table III, this algorithm presented the lowest percentage of error (MAPE) when predicting the exchange rate from US dollars (USD) to euros (EUR). As for the worst result, it occurred when predicting the exchange rate from euros (EUR) to Colombian pesos (COP) with an error of 0.67%.

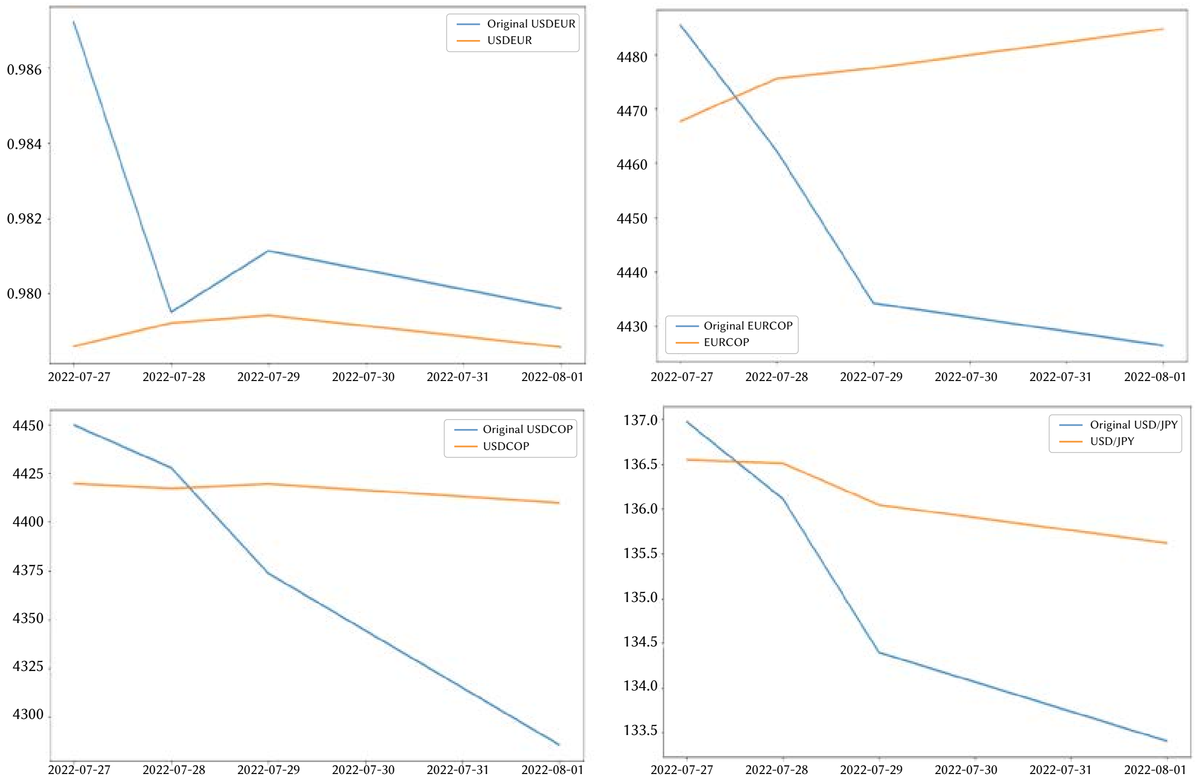


Fig. 15. Actual vs estimated value chart for USD/EUR, USD/JPY, USD/COP and EUR/COP using BCVAR algorithm with h=4 (days ahead).

TABLE III. Comparative Table SVR-WAVELET

Comparative table SVR-WAVELET				
	USD/EUR	USD/JPY	USD/COP	EUR/COP
MAE	0.0029	0.4850	26.7700	28.3900
RMSE	0.0039	0.7460	37.1700	37.0700
MAPE	0.30%	0.35%	0.65%	0.67%

TABLE IV. Comparative Table BCVAR

Comparative table BCVAR				
	USD/EUR	USD/JPY	USD/COP	EUR/COP
MAE	0.0007	1.8674	82.0599	11.0077
RMSE	0.0007	1.8674	82.0599	11.0077
MAPE	0.07%	1.40%	1.91%	0.25%

2. Performance: BCVAR Model

The BCVAR algorithm, described above, works when there are many variables within the system to be predicted and it becomes computationally difficult to find the relationship between them. The logic behind the algorithm is to randomly generate compression matrices and then apply the VAR model to predict the value of the currency "h" days in the future, using these matrices along with the input variables.

In this particular case, to predict the value of USD/EUR, USD/JPY, USD/COP and EUR/COP, other currencies were included in the input matrix in order to determine their possible influence on the value of the target currency. These additional currencies are USD/CAD, USD/AUD, EUR/AUD, EUR/CAD, CAD/JPY, EUR/JPY and EUR/MXN, which correspond to the Canadian dollar, the Australian dollar, the Australian euro, the Canadian euro, the Canadian dollar against the Japanese yen, the euro against the Japanese yen and the euro against the Mexican peso, respectively. The prediction was made by days, and based on criteria defined by h=1, 4 and 12 days in the future, based on criteria defined in the model by T. Paponpat [16].

Performing the estimate with 1 day in the future (h=1), the results shown in Table IV are found.

An estimate was made with 4 days in the future (h=4), obtaining the results shown in Fig. 15.

3. Resume of All Algorithms

Finally, Table V presents a summary of the comparison of the five algorithms implemented using as a case study the results obtained with US/COP currencies.

TABLE V. Chart of US/COP Currency Results

	MAE	MAPE	RMSE
BCVAR	113.96	2.43%	134.43
RBFNN	10.05	0.22%	11.20
ASTWITTER	151.06	3.27%	186.71
LSTM+CNN	58.49	1.24%	71.62
WAVELETS+SVM	55.93	1.18%	73.66

4. Currency Comparison US/COP

According to the results obtained, it can be indicated that the algorithm that uses radial functions (RBF) to improve the training of a classical neural network is the one that shows the best performance in terms of accuracy. On the other hand, it was found that the algorithm that combines sentiment analysis on Twitter information with support vector regression (SVR) obtained the worst performance. Although the incorporation of natural language processing to the analysis can enrich the input information to the model, it is considered that the substitution of the SVR technique by a neural network could improve the performance of this algorithm. Finally, the RBF NN algorithm has the best performance. Sentiment analysis with SVM is not as accurate. BCVAR is complex, however it is not that precise.

VII. Conclusions

The following article proposed the design of a platform through the use of algorithms based on Machine Learning for currency prediction, in a timely manner. A software architecture was defined that allows to address the problem of currency prediction by incorporating several algorithms identified within the literature review, with the purpose of comparing different techniques and algorithms used to make a comparison of their performance. The results of this proposal can be used in the context of those components and requirements necessary in a computer equipment that allows the development of a platform and the incorporation of new prediction models. Another interesting feature of the developed prototype is its ability to be configured to use other currencies from the foreign exchange market, as well as the stock market. This is because the input defined for each model is a time series and the platform can communicate with Yahoo Finance. It would only be necessary to add a list of the other currencies or stock prices to be consulted and pass them as inputs to the models. This demonstrates the versatility and scalability of the prototype, making it a useful and adaptable tool for predicting different financial markets. For the approach of the architecture, the use of microservices was chosen, which allows to have a modularized system in small services, as well as facilitates the maintenance of each of them by different development teams. In addition, since each service is independent, if one of them fails, the entire system will not be affected and will continue to function with the components that are working properly. Finally, the implemented algorithms have the ability to complement each other according to the needs and currency scenarios.

The literature review found that the most common metrics used by authors to measure the performance of different algorithms and make comparisons are: MAE (Mean Absolute Error), RMSE (Root Mean Square Error) and MAPE (Mean Absolute Percentage Error). These prediction error measures provide a good statistical idea of how far the estimates are from the real values, which allows comparing the performance of algorithms developed or improved by other authors. Other models, although they also use complex mathematical elements such as BCVAR, did not present results as promising as the RBF NN algorithm. It is possible that BCVAR is optimized for stock price prediction rather than for currency price calculation, which could explain the results obtained.

As critical implications, three challenges were identified in the design of the platform. First, the integration of multiple algorithms using different technologies and libraries, which was addressed by container virtualization. Second, because many of these algorithms require a prior training stage that can last several minutes, it was decided to decouple the system into small services that communicate asynchronously and report their status through an event manager. While this decoupled design solves the problem of high processing times, it also introduces a small delay in the responses perceived by the user, which is not critical in this case since it is a prediction system.

As future work, modules can be proposed that allow the combination of compatible algorithms and the introduction of new measurement parameters, in order to improve the performance of the predictions in accordance with the foreign exchange market to be worked on. The platform has the ability to be extended to include more data sources as needed, meaning it could be used to make predictions on other time series, such as stock price in the stock markets. In future iterations, a module could be added to incorporate additional data sources to those described in this document.

References

[1] A. Hernandez-Aguila, M. García-Valdez, J. -J. Merelo-Guervós, M. Castañón-Puga and O. C. López, “Using Fuzzy Inference Systems for the Creation of Forex Market Predictive Models,” IEEE Access, vol. 9, pp. 69391-69404, 2021, doi: 10.1109/ACCESS.2021.3077910.

[2] Md. S. Islam, E. Hossain, A. Rahman, M. S. Hossain, and K. Andersson, “A review on recent advancements in FOREX currency prediction,” Algorithms, vol. 13, no. 8, p. 186, Jul. 2020, doi: 10.3390/a13080186.

[3] S. W. Sidehabi, Indrabayu and S. Tandungan, “Statistical and Machine Learning approach in forex prediction based on empirical data,” 2016 International Conference on Computational Intelligence and Cybernetics, Makassar, Indonesia, 2016, pp. 63-68, doi: 10.1109/CyberneticsCom.2016.7892568.

[4] S. R. Das, D. Mishra, y M. Rout, “A hybridized ELM-Jaya forecasting model for currency exchange prediction”, Journal of King Saud University - Computer and Information Sciences, vol. 32, núm. 3, pp. 345–366, 2020. doi: 10.1016/j.jksuci.2017.09.006.

[5] D. Pradeepkumar and V. Ravi, “Soft computing hybrids for FOREX rate prediction: A comprehensive review,” Computers & Operations Research, vol. 99, pp. 262–284, May 2018, doi: 10.1016/j.cor.2018.05.020.

[6] N. Sung et al., “NSML: a machine learning platform that enables you to focus on your models,” arXiv (Cornell University), Jan. 2017, doi: 10.48550/arxiv.1712.05902.

[7] R. J. Borgli, H. K. Stensland, P. Halvorsen and M. A. Riegler, “Saga: An Open Source Platform for Training Machine Learning Models and Community-driven Sharing of Techniques,” 2019 International Conference on Content-Based Multimedia Indexing (CBMI), Dublin, Ireland, 2019, pp. 1-4, doi: 10.1109/CBMI.2019.8877455.

[8] Z S. Zhao, M. Talasila, G. Jacobson, C. Borcea, S. A. Aftab and J. F. Murray, “Packaging and Sharing Machine Learning Models via the Acumos AI Open Platform,” 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Orlando, FL, USA, 2018, pp. 841-846, doi: 10.1109/ICMLA.2018.00135.

[9] H. N. Semiromi, S. Lessmann, and W. Peters, “News will tell: Forecasting foreign exchange rates based on news story events in the economy calendar,” The North American Journal of Economics and Finance, vol. 52, p. 101181, Mar. 2020, doi: 10.1016/j.najef.2020.101181.

[10] C.-H. Chen, P.-Y. Chen, J. Chun-Wei Lin, “An Ensemble Classifier for Stock Trend Prediction Using Sentence-Level Chinese News Sentiment and Technical Indicators,” International Journal of Interactive Multimedia and Artificial Intelligence, vol. 7, no. 3, pp. 53-64, 2022, doi: 10.9781/ijimai.2022.02.004.

[11] A. Andueza, M. Á. Del Arco-Osuna, B. Fornés, R. González-Crespo, J.-M. Martín-Álvarez, “Using the Statistical Machine Learning Models ARIMA and SARIMA to Measure the Impact of Covid-19 on Official Provincial Sales of Cigarettes in Spain”, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 8, no. 1, pp. 73-87, 2023, doi: 10.9781/ijimai.2023.02.010.

[12] H. Alaidaros, M. Omar, and R. Romli, “The state of the art of agile kanban method: challenges and opportunities,” Independent Journal of Management & Production, vol. 12, no. 8, pp. 2535–2550, Dec. 2021, doi: 10.14807/ijmp.v12i8.1482.

[13] M. Bansal, A. Goyal, and A. Choudhary, “Stock Market Prediction with High Accuracy using Machine Learning Techniques,” Procedia Computer Science, vol. 215, pp. 247–265, Jan. 2022, doi: 10.1016/j.procs.2022.12.028.

[14] M. Biswas, A. Shome, M.A. Islam, A.J. Nova, S. Ahmed, “Predicting Stock Market Price: A Logical Strategy using Deep Learning,” in 2021 IEEE 11th IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE) (pp. 218-223). Penang, Malaysia: IEEE, 2021, doi: 10.1109/ISCAIE51753.2021.9431817.

[15] M. S. Raimundo and J. Okamoto, “SVR-wavelet adaptive model for forecasting financial time series,” 2018 International Conference on Information and Computer Technologies (ICICT), DeKalb, IL, USA, 2018, pp. 111-114, doi: 10.1109/INFOCT.2018.8356851.

[16] P. Taveeapiradeecharoen, K. Chamnongthai and N. Aunsri, “Bayesian Compressed Vector Autoregression for Financial Time-Series Analysis and Forecasting,” IEEE Access, vol. 7, pp. 16777-16786, 2019, doi: 10.1109/ACCESS.2019.2895022.

[17] G. Koop, D. Korobilis, and D. Pettenuzzo, “Bayesian compressed vector

autoregressions," Journal of Econometrics, vol. 210, no. 1, pp. 135-154, 2018, doi: 10.1016/j.jeconom.2018.11.009.

[18] I. E. Livieris, E. Pintelas, and P. Pintelas, "A CNN–LSTM model for gold price time-series forecasting," Neural Computing and Applications, vol. 32, no. 23, pp. 17351–17360, 2020, doi: 10.1007/s00521-020-04867-x.

[19] H. Fan, M. Jiang, L. Xu, H. Zhu, J. Cheng, and J. Jiang, "Comparison of long short term memory networks and the hydrological model in runoff simulation," Water, vol. 12, no. 1, p. 175, 2020, doi: 10.3390/w12010175.

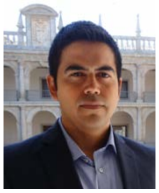
[20] J. Kordonis, S. Symeonidis, and A. Arampatzis, "Stock price forecasting via sentiment analysis on twitter", in Proceedings of the 20th Pan-Hellenic Conference on Informatics, 2016, pp. 1–6. doi: 10.1145/3003733.3003787.

[21] S. Yu and J. Ou, "Forecasting Model of Agricultural Products Prices in Wholesale Markets Based on Combined BP Neural Network -Time Series Model," 2009 International Conference on Information Management, Innovation Management and Industrial Engineering, Xi'an, China, 2009, pp. 558-561, doi: 10.1109/ICIII.2009.140.

[22] A. Botchkarev, "A New Typology Design of Performance Metrics to Measure Errors in Machine Learning Regression Algorithms", Interdisciplinary Journal of Information, Knowledge, and Management, vol. 14, pp. 45-76, 2019, doi: 10.28945/4184.

Kevin Gordillo Orjuela

He has been employed as a software engineer in the private sector since 2016 and has been an active member of the GIIRA research group since 2019. Holding a Master's degree in Information Science and Communication from Universidad Distrital Francisco José de Caldas (Bogotá, Colombia - 2023), he also earned a Bachelor's degree in Systems Engineering from the same university in 2017. His research interests span various domains including web science, machine learning, data analysis, software architecture, information visualization, and visual analytics.



Paulo Alonso Gaona García

He earned a Ph.D. in Information and Knowledge Engineering from University of Alcalá in 2014 and working as a collaborator researcher in the Information Engineering Research Unit at University of Alcalá since 2012. Is full professor at Engineering Faculty of Universidad Distrital Francisco José de Caldas, Bogotá – Colombia since 2008. He is director of Multimedia Research Group and active member of GIIRA research group since 2008. He has a Master in Information Science and Communication from Universidad Distrital Francisco José de Caldas (Bogotá - Colombia - 2006). He is Systems Engineer at Universidad Distrital Francisco José de Caldas (2003). His research interest includes web science, semantic web, network and communications, e-learning, information visualization and visual analytics.

Carlos Enrique Montenegro Marin

PhD in Computer Science from Oviedo University (2012). Master in Web Site Management and Engineering at the International University of the Rioja - UNIR (2013). Master in Information Science at the Universidad Distrital Francisco José de Caldas (2006). Systems Engineer at the District University (2003). He was a Dean of Engineering Faculty (December 2018 - 2019) and full professor, attached to Engineering Faculty of District University "Francisco José de Caldas" since 2006. He was Coordinator of committee of accreditation in the bachelor's in systems engineering (2007 to 2010), He was Coordinator of the bachelor's in systems engineering (2012 to 2014).

Advances in AI-Generated Images and Videos

Hessen Bougueffa[1], Mamadou Keita[1], Wassim Hamidouche[2], Abdelmalik Taleb-Ahmed[1], Helena Liz-López[3], Alejandro Martín[3], David Camacho[3], Abdenour Hadid[4] *

[1] Laboratory of IEMN, CNRS, Centrale Lille, UMR 8520, Univ. Polytechnique Hauts-de-France (France)
[2] Univ. Rennes, INSA Rennes, CNRS, IETR - UMR, Rennes, 6164 (France)
[3] Computer Systems Department, Universidad Politécnica de Madrid (Spain)
[4] Sorbonne Center for Artificial Intelligence, Sorbonne University Abu Dhabi (United Arab Emirates)

* Corresponding author: bougueffaeutamenehessen@gmail.com (H. Bougueffa), Mamadou.Keita@uphf.fr (M. Keita), whamidouche@gmail.com (W. Hamidouche), abdelmalik.taleb-ahmed@uphf.fr (A. Taleb-Ahmed), helena.liz@upm.es (H. Liz-López), alejandro.martin@upm.es (A. Martín), david.camacho@upm.es (D. Camacho), abdenour.hadid@ieee.org (A. Hadid).

Received 13 November 2024 | Accepted 20 November 2024 | Early Access 28 November 2024

Abstract

In recent years generative AI models and tools have experienced a significant increase, especially techniques to generate synthetic multimedia content, such as images or videos. These methodologies present a wide range of possibilities; however, they can also present several risks that should be taken into account. In this survey we describe in detail different techniques for generating synthetic multimedia content, and we also analyse the most recent techniques for their detection. In order to achieve these objectives, a key aspect is the availability of datasets, so we have also described the main datasets available in the state of the art. Finally, from our analysis we have extracted the main trends for the future, such as transparency and interpretability, the generation of multimodal multimedia content, the robustness of models and the increased use of diffusion models. We find a roadmap of deep challenges, including temporal consistency, computation requirements, generalizability, ethical aspects, and constant adaptation.

Keywords

AI-Generated Content, Image Generation, Multimodal, Video Generation.

DOI: 10.9781/ijimai.2024.11.003

I. Introduction

The recent progress in Artificial intelligence (AI) has led to a revolution in the creation of synthetic images and videos, mainly due to the remarkable capabilities of advanced generative models, diffusion models, or Generative adversarial networks (GANs), among others. There are now a large number of applications and tools available to users, such as DALL-E [1], GLIDE [2], Midjourney [3], Imagen [4], VideoPoet [5], Sora [6], or Genie [7]. These tools are designed to produce realistic and believable digital content easily. This development has had a profound impact, with various applications across different areas.

These techniques are capable of generating multimedia content on any topic or object. Therefore, there are countless opportunities, especially in some application domains, which can benefit greatly from these techniques and tools: *entertainment and media*, allowing the generation of characters, scenarios or elements that would be very difficult to create by traditional means [8]–[10]; *creative industries*, allows artists to streamline their work and improve its quality, for example by creating sketches to work on further, or creating elements to add to their work [11], [12]; *education*, creating engaging educational content, including simulations and visual aids to help illustrate and clarify complex ideas, and adapting to different learning styles [13], [14]; *security and forensics*, helping to create robust models capable of detecting false or generated information more easily, for example by assisting in data augmentation [15], [16]. As we can see, the applications of these techniques are limitless, and as their capabilities improve, they can be more easily applied to different problems in society.

This collection of tools and methodologies not only presents advantages, but also a number of weaknesses and potential risks that need to be carefully analysed. The ability to produce highly realistic synthetic media easily causes concern about their possible inappropriate use. Deepfakes and other kinds of manipulated content can be used to spread misinformation, create disinformation, and manipulate public opinion, undermining trust in digital media [17], [18]. This dual potential for both positive and negative impact highlights a crucial problem. While leveraging the benefits of generative models, there is an urgent need to develop effective detection methods to distinguish between real and AI-generated content. As generative models become more sophisticated, the task of detecting synthetic media becomes increasingly complex, necessitating the continuous evolution of detection techniques.

Please cite this article as:
H. Bougueffa, M. Keita, W. Hamidouche, A. Taleb-Ahmed, H. Liz-López, A. Martín, D. Camacho, A. Hadid. Advances in AI-Generated Images and Videos, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 1, pp. 173-208, 2024, http://dx.doi.org/10.9781/ijimai.2024.11.003

Despite the significant advancements in generative models, several gaps and challenges persist in both their deployment and the methods used to detect synthetic media. One major challenge lies in the resource-intensive nature of training and deploying these models. High computational requirements limit accessibility, particularly for smaller organizations and researchers lacking the necessary infrastructure to fully utilise these technologies. This creates a barrier to wider adoption and raises concerns about the scalability and sustainability of generative models as they continue to evolve. Furthermore, even advanced models such as GLIDE [2] and DALL·E 2 [1] encounter challenges when processing complex prompts. These challenges can limit their ability to generate high-quality outputs under specific conditions. Similarly, Imagen [19] enhances computational efficiency but still grapples with resource demands and complex prompts. These limitations underscore a need for improved flexibility and robustness in current generative technologies.

On the video generation front, text-to-video models face significant challenges in maintaining high fidelity and continuity of motion over extended sequences. Many existing methods simply extend text-to-image models, which do not fully address the unique complexities inherent in video generation. This highlights the need for more specialized approaches that can effectively handle the temporal dynamics and continuity required for high-quality video content.

Detecting synthetic media presents significant challenges. Current detection models struggle to keep pace with the rapid advancements in generative technologies, making it difficult to reliably differentiate between real and AI-generated images and videos. These models tend to specialize in the types of synthetic content they were trained on, leading to poor performance when faced with new data from different or updated models. Additionally, detection algorithms must be resilient against various transformations and adversarial attacks [20], [21], such as image compression and blurring, which can significantly diminish their effectiveness. Techniques for identifying deepfakes [22] and other forms of image and video forgeries [23] also encounter obstacles due to the constantly evolving nature of these manipulations and the need for high-quality datasets and standardized benchmarks.

To address these challenges and advance the field, this survey:

- Presents an updated picture of synthetic image generation and detection techniques.
- Presents an overview of video generation and detection techniques.
- Provides a list of the main video and image datasets used by researchers.
- Describes trends, challenges and research directions that can be explored in the AI generation, in video and image, and supports them with the conclusions of the analysis.

By providing a thorough examination of both the generative and detection aspects of synthetic media, this survey aims to foster a deeper understanding of the current challenges and opportunities in the field, promoting the development of technologies that can maximize the benefits of AI-generated content while minimizing its risks.

This survey is structured to comprehensively address both the generative capabilities and detection techniques of AI-generated images and videos, see Fig. 1. Section II reviews related works and surveys, providing a foundation for understanding the current state of research in this domain. Section III dives into image generation and detection, detailing various advanced generative models and the methods used to detect synthetic images. Section IV focuses on video generation and detection, exploring the advancements in video generation and the techniques to identify AI-generated videos. Section V discusses the datasets used for generative and detection algorithms, highlighting the importance of diverse and high-quality datasets. Section VI identifies the ongoing challenges in both generating and detecting synthetic media. Finally, Section VII concludes the survey by summarizing the key findings and suggesting future directions for research and development in this field.

II. Related Work and Related Surveys

The field of AI-generated images and videos has been extensively studied, with several surveys reviewing the advancements and challenges in this area. This section provides an overview of key surveys and positions our work in relation to them, highlighting the unique aspects of our approach, summarised in Table I.

- Liu *et al.* [24] conducted an extensive review on human image generation, categorizing existing techniques into three main paradigms: data-driven, knowledge-guided, and hybrid. The survey covers the most representative models and approaches within each paradigm, highlighting their specific advantages and limitations. Additionally, it explores a range of applications, datasets, and evaluation metrics relevant to human image generation. The paper also addresses the challenges and potential future directions in the field, offering valuable insights for researchers interested in this rapidly evolving domain.
- Chen *et al.* [28] concentrated on controllable text-to-image generation models. They investigated various methods that precisely control the produced content, such as personalized and multi-condition generation techniques. The authors explore the practical applications of these models in content creation and design while also recognizing current constraints and suggesting future directions to enhance the adaptability and accuracy of these generative models.
- Joshi *et al.* [29] provided an extensive analysis on the use of synthetic data in human analysis, focusing on the advantages and challenges in biometric recognition, action recognition, and person re-identification. The survey delves into various techniques for generating synthetic data, including deep generative models and 3D rendering tools, emphasizing their potential to tackle issues related to data scarcity, privacy concerns, and demographic biases in training datasets. Additionally, the authors explore how synthetic data can augment real datasets to enhance model performance scalability analysis and simulate complex scenarios that are challenging to capture with real data. They also address concerns about synthetic datasets, such as identity leakage and lack of diversity.
- Figueira *et al.* [25] focused on the generation of synthetic data with Generative Adversarial Networks (GANs). The authors emphasize the significance of synthetic data, particularly in cases where data is limited or contains sensitive information. They highlight how GANs can proficiently create high-quality synthetic samples that imitate real data distributions. This study presents a detailed summary of current methods and challenges in synthetic data generation, emphasizing the utilization of GANs for diverse data types, including tabular data, and exploring various GAN architectures that cater to these requirements.
- Nguyen *et al.* [26] offered a comprehensive review of deepfake generation and detection methods using deep learning techniques. They explored different types of deepfakes, such as face-swaps, lip-syncs, and puppet-master variations, while highlighting the progress and challenges in identifying these manipulations. The survey covers traditional and deep learning-based approaches for detecting deepfakes, including methods based on manual feature creation and those utilizing deep neural networks. Their work emphasizes the importance of developing robust detection algorithms to counter

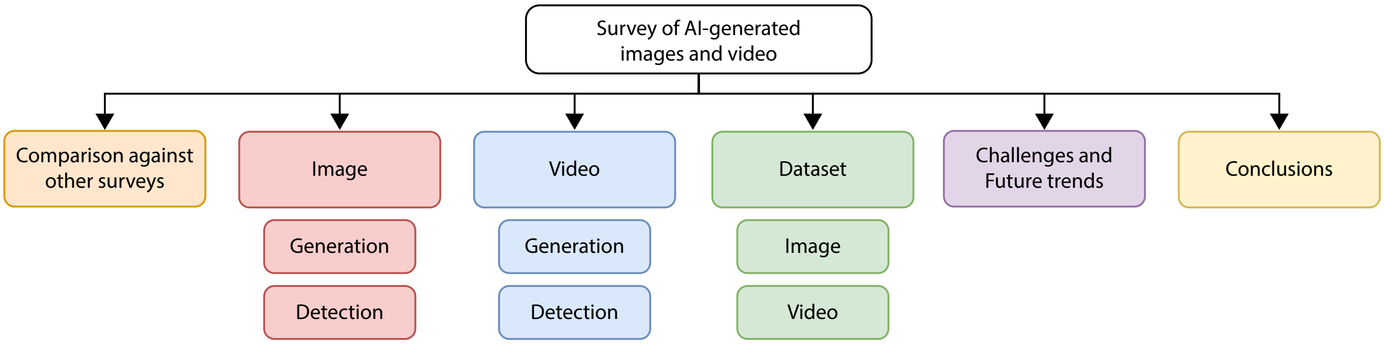

Survey of AI-generated images and video
Comparison against other surveys
Image
Generation
Detection
Video
Generation
Detection
Dataset
Image
Video
Challenges and Future trends
Conclusions

Fig. 1. Schematic representation of the structure followed.

TABLE I. Comparison of Previous Literature Reviews

Authors	Year	Task analysed		Modalities		Main Contribution	Limitations
		Generation	Detection	Image	Video		
Liu *et al.* [24]	2022	✓	×	✓	✓	It provided an extensive review on the generation of human images	It only deals with the generation of human images, without covering other possible scenarios.
Zhang *et al.* [22]	2022	✓	✓			It provides a detailed analysis of video and image sample manipulation and detection techniques.	Focus on the manipulation of video and image samples.
Figueira *et al.* [25]	2022	✓	×	✓	×	It provides a very detailed analysis of the use of GANs within data generation, focusing on training problems and evaluation techniques.	It does not focus on image and video generation.
Nguyen *et al.* [26]	2022	✓	✓	✓	✓	It analyses both the techniques of generation, or manipulation, and the detection of images and videos.	It is mainly focused on the manipulation of multimedia data, not so much on the generation of synthetic samples.
Tyagi *et al.* [23]	2023	✓	✓	✓	✓	Performs a detailed analysis of manipulation and detection techniques for video and audio samples.	The focus is not on synthetic sample generation and detection techniques, but on manipulation techniques.
Bauer *et al.* [27]	2024	✓	×	✓	✓	It performs one of the most comprehensive data generation analyses available.	It is not focused on the generation of image and video samples.
Chen *et al.* [28]	2024	✓	×	✓	×	It covers one of the newest approaches to image generation, diffusion models for Text-to-image task.	This is a very limited survey, as it covers only one of the imaging approaches, without analysing other techniques or modalities.
Joshi *et al.* [29]	2024	✓	×	✓	✓	Explores techniques including improving model performance, increasing data diversity and scalability, and mitigating privacy issues.	It only focuses on generating samples that represent humans, leaving a large part of the field unstudied.

the increasing complexity of deepfake creation techniques. This study holds particular relevance in developing new multimodal approaches for deepfake detection, which are in alignment with investigating cross-modality fusion strategies.

- Bauer *et al.* [27] examined Synthetic Data Generation (SDG) models, analyzing 417 models developed over the past decade. The survey classifies these models into 20 distinct types and 42 subtypes, providing a comprehensive overview of their functions and applications. The authors identified significant model performance and complexity trends, highlighting the prevalence of neural network-based approaches in most domains, except privacy-preserving data generation. The survey also discusses challenges, such as the absence of standardized evaluation metrics and datasets, indicating the need for enhanced comparative frameworks in future research.
- Zhang *et al.* [22] analysed the generation and detection of deepfakes, shedding light on both the progress made and the challenges encountered in this area. They outline two main techniques for creating deepfakes, face swapping and facial reenactment, and discuss the impact of GANs and other deep learning methods. Their work also explores various detection strategies, ranging from biometric and model features to machine learning-based methods. They emphasize the persistent challenges arising from evolving deepfake technologies, the need for high-quality datasets, and the absence of a standardized benchmark for detection methods. This survey is essential for gaining insights into the current state of generating and detecting deepfakes, which present significant challenges to privacy, security, and societal trust.
- Tyagi *et al.* [23] conducted a comprehensive analysis of image and video forgery detection techniques, highlighting the various manipulation methods, such as morphing, splicing, and retouching, and the challenges associated with detecting these alterations in digital media. The survey also reviewed different datasets used for training and evaluating forgery detection algorithms, emphasizing the need for robust, generalized methods capable of detecting multiple types of manipulations across diverse visual datasets. This work provides a detailed examination of both traditional and deep learning-based approaches, illustrating the advancements and limitations in the field of digital media forensics.

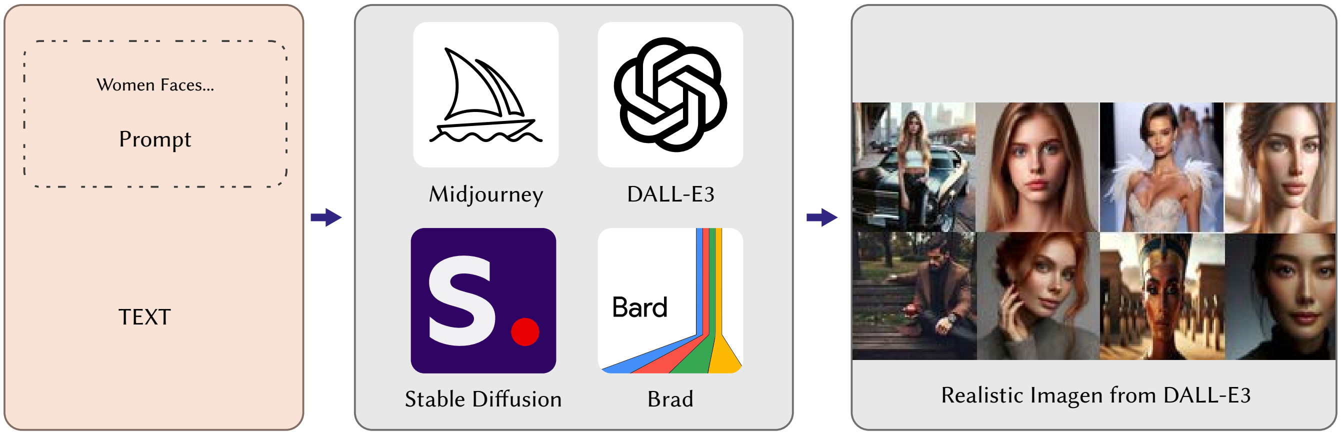


Fig. 2. Overview of the main approaches to image generation with AI.

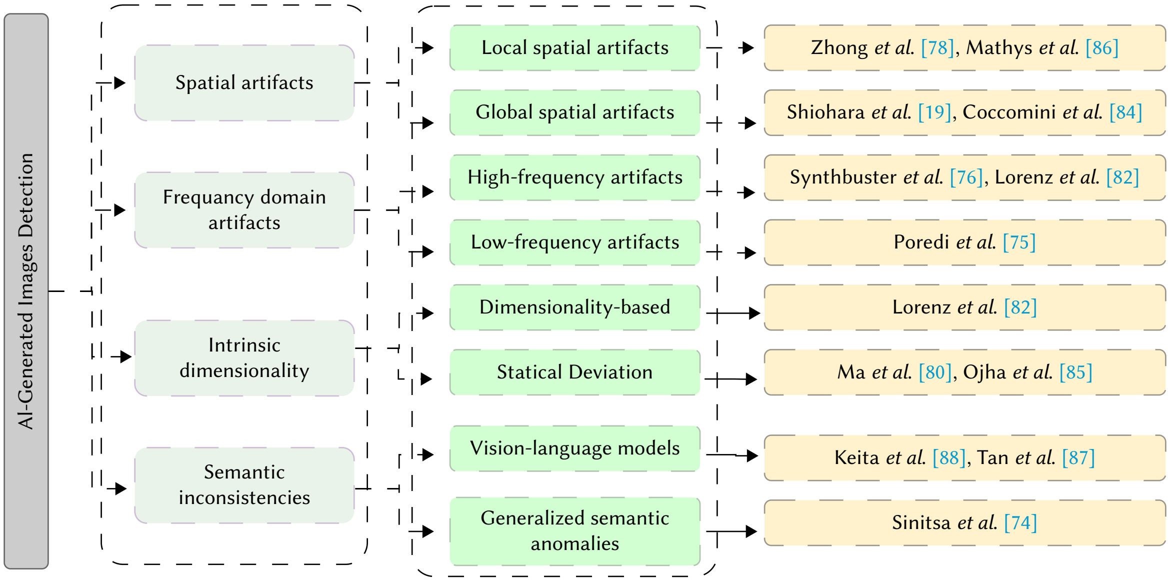

AI-Generated Images Detection
Spatial artifacts
Frequancy domain artifacts
Intrinsic dimensionality
Semantic inconsistencies
Local spatial artifacts
Global spatial artifacts
High-frequency artifacts
Low-frequency artifacts
Dimensionality-based
Statical Deviation
Vision-language models
Generalized semantic anomalies
Zhong et al. [78], Mathys et al. [86]
Shiohara et al. [19], Coccomini et al. [84]
Synthbuster et al. [76], Lorenz et al. [82]
Poredi et al. [75]
Lorenz et al. [82]
Ma et al. [80], Ojha et al. [85]
Keita et al. [88], Tan et al. [87]
Sinitsa et al. [74]

Fig. 3. Overview of AI-generated Image Detection.

As we can see, this survey has a number of advantages over other published reviews of the field. Firstly, it is the first work to focus exclusively on synthetic sample generation techniques, which also provides a list of datasets published in recent years. It also analyses the approaches with which researchers are tackling the problem of detecting these synthetic samples.

III. AI Image Generation and Detection

In this section, we will focus on the generation of images with AI techniques, as well as on the main approaches for their detection. As mentioned above, AI, more specifically Deep Learning (DL) has shown significant progress in the fields of image generation and detection.

Advanced models have greatly improved the ability to **generate synthetic images**, focusing on enhancing aspects such as image quality and realism. Recent developments have led to improved training stability and higher-quality generated images, addressing common challenges and allowing for the creation of diverse and realistic outputs. Innovations in model architectures have also provided greater control over the image generation process, resulting in even more varied and convincing synthetic images. Fig. 2 illustrates a subset of AI-generated image and video techniques, specifically focusing on generative models that rely on text or prompts to create the samples. While this figure highlights key models used in text-to-image or text-to-video synthesis, other generative approaches are discussed in the subsequent sections.

Models for **synthetic images detection** have also made substantial progress. These detection models have become more advanced, using deep learning techniques to identify subtle artifacts and inconsistencies in generated images. As a result, they are crucial in differentiating between real and synthetic images, ensuring the integrity of visual content. The ongoing evolution of these models indicates the dynamic nature of the field, with continuous research efforts focused on improving their precision and resilience [30], [31].

A. Image Generation

Within AI image generation, we will analyse two different approaches, see Fig. 3. The first approach, ***Text-to-image synthesis***, will focus on generating image samples from text descriptions; while the second approach, ***Image-to-image translation***, focuses on modifying the original image while preserving some visual properties in the final sample. A concise summary of the main image generation techniques is presented in Table II.

TABLE II. Comprehensive Overview of a Few Synthetic Image Generation Techniques

Models	Year	Technique	Target Outcome	Data Used	Open Source
NVAE [66]	2020	Hierarchical VAE	High-fidelity images	CelebA, FFHQ	No
CogView [41]	2021	Transformer-based	Text-to-image synthesis	Diverse text and images	Yes
StyleGAN3 [59]	2021	GAN-based	High-quality images	FFHQ, CelebA	Yes
BigGAN [73]	2021	GAN-based	Large-scale image synthesis	ImageNet	Yes
GLIDE [2]	2021	Diffusion-based	Generate images from text prompts	DALL-E's dataset	Yes
DALL-E 2 [1]	2022	Transformer-based	Text-to-image synthesis	Custom, diverse content	Yes
DiVAE [38]	2022	VQ-VAE with diffusion	High-quality reconstruction	ImageNet	No
VQ-VAE-2 [65]	2022	VA E -based	High-resolution images	Large-scale datasets	Yes
EfficientGAN [61]	2022	GAN-based	Efficiency and quality	Custom datasets	Partial
Latent Diffusion [43]	2023	Diffusion-based	Photorealistic images	Various	Yes
DALL-E 3 [51]	2023	Enhanced Transformer	Improved prompt following	Custom image captioner dataset	No
Imagen [4]	2023	Transformer-based	High-fidelity image synthesis	Open Images, ImageNet	No
Imagen2 [50]	2023	Style-conditioned diffusion	Lifelike images with context	Diverse dataset	No
Muse [40]	2023	Transformer T5-XXL	High-fidelity zero-shot editing	CC3M, COCO	No
SDXL [48]	2023	Stable Diffusion	High-resolution image synthesis	Custom dataset	Yes
StyleGAN-T [32]	2023	GAN-based	High-quality image synthesis	Comprehensive dataset with various text-image pairs	Yes
GALIP [35]	2023	GAN-based, utilizing CLIP	Efficient quality image creation from text	Diverse datasets	Yes
GigaGAN [33]	2023	Advanced GAN	High-resolution, detailed image generation from text	Extensive datasets with diverse image-text pairs	Yes
UFOGen [37]	2024	GAN and diffusion	High-quality fast generation	-	No
RAPHAEL [49]	2024	Diffusion with MoEs	Artistic images from text	Subset of LAION-5B	Yes
Ahmed et al. [36]	2024	GAN with spatial co-attention	Enhanced image generation	CUB, Oxford-102, COCO	No

1. Text-to-Image Synthesis

In this section, we will look at different approaches to creating synthetic images from text. As this is a growing field we can observe a variety of different techniques, such as GANs, transformers or diffusion models.

Generative Adversarial Networks: Some authors continue to focus on GANs which, although not particularly novel, have competitive results in the field. For example, Sauer *et al.* [32] have improved the robust StyleGAN architecture to develop StyleGAN-T. This model tackles the challenge of producing visually diverse and attractive images from textual descriptions at scale, effectively speeding up the process while maintaining image fidelity. StyleGAN-T is trained on a comprehensive dataset containing various text-image pairs, ensuring diverse visual outputs. However, one limitation is the potential for reduced accuracy in rendering complex scenes due to the inherent challenges of text ambiguity and the current limitations of GANs in understanding nuanced textual descriptions. Kang *et al.* [33] proposed GigaGAN's, an architecture that includes an improved generator and discriminator that efficiently handle large-scale data, allowing for the creation of diverse and visually compelling images. However, like other large-scale GANs, GigaGAN requires significant computational resources for training and has the potential to overfit precise textual descriptions if the training data lacks diversity. Despite these limitations, GigaGAN's image synthesis capability is a powerful tool in AI-driven creative image generation, expanding the boundaries of machine understanding and visualization of textual content. The model TextControlGAN [34] introduces an innovative method to improve text-to-image synthesis by modifying the Generative Adversarial Network (GAN) architecture. This modification aims to enhance control and precision in generating images from textual descriptions, by integrating specific control mechanisms within the GAN framework. This capability is essential for applications that require high fidelity between textual inputs and visual outputs, such as in digital media creation and automated content generation.

Other authors have explored the option of **a combination between GAN with other types of techniques** such as the CLIP model, such as Ming Tao *et al.* [35] have applied the pre-trained CLIP model to Generative Adversarial Networks (GANs) to transform the process of text-to-image synthesis. This innovative approach enhances the efficiency and quality of the images created from textual descriptions. By integrating CLIP into both the discriminator and generator, the model achieves strong scene understanding and domain generalization using fewer parameters and less training data. By leveraging diverse and extensive datasets, this method enables the generation of a broad range of intricate and visually appealing images. This approach accelerates the synthesis process and ensures a smoother and more controllable latent space, thereby significantly reducing the computational resources typically required for high-quality image synthesis.

Ahmed *et al.* [36] proposed a novel approach that involves simultaneously generating images and their corresponding foreground-background segmentation masks. This is achieved by using a new Generative Adversarial Network (GAN) architecture named COS-GAN, which incorporates a spatial co-attention mechanism to improve the quality of both the images and segmentation masks. The innovative aspect of COS-GAN lies in its ability to handle multiple image outputs and their segmentations from textual descriptions, thereby enhancing applications such as object localization and image editing. It was extensively tested on diverse datasets, including CUB, Oxford-102, and COCO. However, it faces challenges, such as the high computational demand required for training and potential biases embedded within the large-scale datasets used. These limitations could impact the generalizability and ethical deployment. By contrast, Xu *et al.* [37], chose to combine these GAN with diffusion models. They proposed UFOGen, that offers a novel approach to generating high-quality images from text quickly. Combining elements of Generative Adversarial Networks (GANs) and diffusion models efficiently creates images in a single step, eliminating the need for slower, multi-step processes used by standard diffusion models. UFOGen's training process is greatly improved by utilizing pre-trained diffusion models, which enhances efficiency and reduces training times. However, similar to other generative models, UFOGen also faces limitations. It depends on large-scale datasets that may contain biased or inappropriate content, potentially leading to biased generated images, which raises ethical concerns and affects the fairness and diversity of the output.

Autoencoder models: Another approach we have seen in the generation of images from text is **autoencoder models**. For example, Saharia *et al.* [4] introduced Imagen, a text-to-image model using classifier-free guidance (CFG) and a pre-trained T5-XXL encoder to improve computational efficiency. The model's key innovation is using large language models to enhance image quality and text-image alignment. Imagen generates images starting at 64×64 resolution, then upscales to 256×256 and 1024×1024 using super-resolution models. Despite achieving a strong FID score of 7.27 on COCO, the model faces challenges with dataset biases, high computational demands, and difficulties in generating realistic human images. On the other hand Shi *et al.* [38] developed DiVAE, which combines a VQ-VAE architecture with a denoising diffusion decoder to create highly realistic images, excelling in image reconstruction and text-to-image synthesis tasks. Using a CNN encoder, the model first compresses images into latent embeddings and then reconstructs them into high-quality images through a diffusion-based decoder. Trained on the ImageNet dataset, DiVAE delivers superior performance in terms of FID scores compared to models like VQGAN. However, the diffusion process is computationally intensive, requiring many steps, and the model is restricted by the fixed image size determined by the training data.

Contrastive learning: it has also been shown that this type of learning is a good technique for tackling this type of task using AI models. The CLIP model [39], created by OpenAI, has attracted the attention of a large number of researchers. This model is able to relate images and text by using contrastive learning, training on large multimodal datasets to align visual and linguistic representations in a shared space, allowing tasks such as image generation, search and classification to be performed without the need for specific supervised training. As a result, it is one of the most widely used approaches for researchers to generate synthetic images from text.

Tranformer: We have also analysed different research that has used transformers for the generation of synthetic images. Muse [40] is a Transformer designed for text-to-image generation. It utilizes a pre-trained T5-XXL language model to predict masked image tokens. Trained on 460 million text-image pairs from CC3M and COCO datasets, this model excels in generating high-fidelity images and supports zero-shot editing, such as inpainting and outpainting. Muse's efficiency exceeds that of diffusion and autoregressive models due to its discrete token space and parallel decoding. However, it faces challenges in rendering long phrases, handling high object cardinality, and managing multiple cardinalities in prompts. Ming Ding *et al.* [41] have introduced CogView. This model harnesses a 4-billion-parameter Transformer architecture in combination with a VQ-VAE tokenizer. CogView operates by encoding text into discrete tokens, which the Transformer processes to forecast corresponding visual tokens. These visual tokens are then transformed into high-quality images using the VQ-VAE decoder. CogView underwent training on extensive datasets, incorporating image-text pairs from diverse sources. Despite its remarkable capabilities, CogView does have limitations. The model demands substantial computational resources for training owing to its expansive parameter size. Similar to numerous text-to-image models, it encounters challenges with intricate or ambiguous text prompts, leading to less precise image generation. Additionally, dependence on extensive datasets can introduce biases within the training data, impacting the variety and impartiality of the generated images. CogView2 [42] used a sophisticated Transformer architecture to quickly generate high-quality images from text. The model begins by producing low-resolution images and then progressively refines them using super-resolution modules, ensuring detailed and consistent results. With a foundation built on a 6-billion-parameter Transformer, the model is trained on diverse datasets of text-image pairs, allowing it to handle tasks such as text-to-image generation, image infilling, and captioning in multiple languages. Nevertheless, CogView2 requires substantial computational resources and careful tuning to balance local and global coherence in the generated images.

Diffusion models. This is one of the topics that has attracted the most researchers. Latent Diffusion Models (LDMs) [43] are a major step forward in high-resolution image synthesis, see Fig. 4. They achieve this by using diffusion models within the latent space of pre-trained autoencoders. This reduces the computational requirements typically associated with diffusion models operating in pixel space while maintaining high visual fidelity. Incorporating cross-attention layers within the UNet backbone is a significant advancement in LDMs. It enables the generation of high-quality outputs based on various input conditions, such as text prompts and bounding boxes. This architecture supports high-resolution synthesis using a convolutional approach. The model is trained to predict a less noisy version of the latent variable by focusing on essential semantic features rather than on high-frequency details that are often imperceptible.

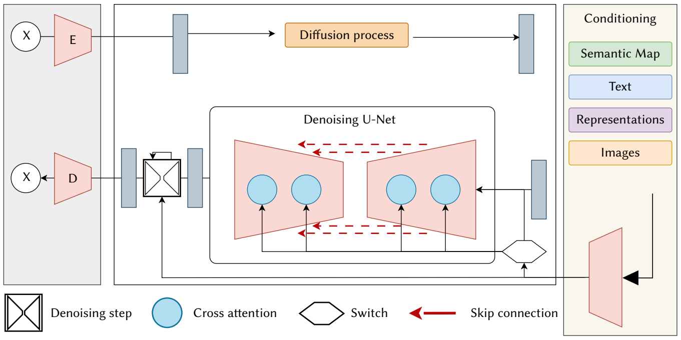


Fig. 4. Latent Diffusion Models architecture from Rombach *et al.* [43].

Anton *et al.* [44] present a new method for synthesizing images from a text by combining image-prior models with latent diffusion techniques. The model utilizes CLIP to map text embeddings to image embeddings and incorporates a modified MoVQ implementation as the image autoencoder. After training on the COCO-30K dataset, Kandinsky achieves high-quality image generation with a competitive FID score. Despite the need for further improvements in the semantic coherence between text and generated images, Kandinsky's versatility in supporting text-to-image generation, image fusion, and inpainting represents a significant advancement in AI-driven image synthesis. EmoGen [45] marks a significant leap forward in text-to-image models. It centers on producing images that capture distinct emotions, solving the difficulty of linking abstract emotions with visual representations. This model excels at creating images that are semantically clear and resonate emotionally. It accomplishes this by aligning the emotion-specific space with the powerful semantic capabilities of the CLIP model. This alignment is established through a mapping network that interprets abstract emotions into concrete semantics, guaranteeing that the generated images faithfully reflect the intended emotional tones. The model has undergone training and validation using EmoSet, a comprehensive visual emotion dataset with detailed attribute annotations, aiding in optimizing the model for diverse and emotionally accurate image generation. Despite its advancements, EmoGen faces challenges akin to other generative models, including reliance on potentially biased large datasets and the substantial computational resources needed for training and inference, limiting its accessibility and applicability across different research groups and practical uses.

Latent Diffusion Models (LDMs) also have their limitations. One significant challenge is the use of large-scale, often uncurated datasets, which can introduce biases and ethical concerns. While LDMs are more computationally efficient than traditional pixel-based diffusion models, they still require substantial computational resources for training and inference, which may be prohibitive for smaller research groups. LDMs also struggle with generating realistic images of people, leading to lower preference rates in evaluations. Additionally, these models can reflect societal biases, highlighting the importance of robust bias mitigation strategies and the need for more ethically curated datasets in future research. Hang Li et al. [46] present an innovative approach focusing on the ethical implications of AI-generated content and introduce a self-supervised method for identifying interpretable latent directions within diffusion models. The objective is to mitigate the generation of inappropriate or biased images, thus enhancing control over the generated images and ensuring they align with ethical standards while avoiding perpetuating harmful stereotypes. The model has been trained on diverse datasets, allowing it to handle a broad scope of concepts sensitively and responsibly. However, the extensive reliance on datasets may introduce potential biases, while the high computational demand for processing these datasets presents challenges for accessibility and scalability.

Some researchers have chosen to combine the CLIP model with diffusion models. For example, Nichol *et al.* [2] introduced GLIDE, a text-to-image diffusion model that replaces class labels with text prompts. It uses classifier guidance, with a CLIP model in noisy image space, and classifier-free guidance [47], which integrates text features directly into the diffusion process. GLIDE's 3.5B parameter model encodes text through a transformer to generate high-quality images. While effective in photorealism and caption alignment, GLIDE struggles with complex prompts and requires substantial computational power. Ramesh *et al.* [1] introduced DALL·E 2, a model leveraging CLIP and diffusion techniques for generating realistic images from text descriptions. DALL·E 2 operates in two stages: a prior model creates a CLIP image embedding from text, followed by a diffusion-based decoder that generates the final image. This architecture ensures both diversity and realism in the output. The model's use of CLIP embeddings captures semantic and stylistic nuances, enabling high-quality image generation and manipulation. Although trained on a vast dataset, DALL·E 2 faces challenges with complex prompts and fine-grained attribute accuracy, highlighting areas for further improvement.

Furthermore, Podell *et al.* [48] developed SDXL, which is a major step forward in high-resolution image synthesis, expanding on the foundational work of Stable Diffusion models. It utilizes a significantly larger UNet backbone, about three times larger than its predecessors, with more attention blocks and a larger cross-attention context. This enhanced architecture enables SDXL to tackle complex text-to-image synthesis tasks effectively. Additionally, SDXL incorporates multiple innovative conditioning schemes and is trained on various aspect ratios, enhancing its versatility in producing images of different resolutions and aspect ratios. Firstly, it generates initial 128×128 latents. Then, a specialized high-resolution refinement model is applied to improve these latents to higher resolutions. The SDXL training involved utilising an improved autoencoder from previous Stable Diffusion versions. It exceeded its predecessors in all assessed reconstruction metrics, ensuring improved local and high-frequency details in the generated images. The final training stage included multi-aspect training with different aspect ratios, further boosting the model's capabilities. Despite its progress, SDXL has some limitations. The model's reliance on large-scale datasets can lead to biases and ethical concerns due to potentially inappropriate content such as pornographic images, racist language, and harmful social stereotypes. SDXL also struggles to create realistic images of people, often resulting in lower preference rates. Furthermore, the model perpetuates existing social biases, favouring lighter skin tones. Xue *et al.* [49] presents Raphael, an innovative method for generating images from text. It aims to create highly artistic images that closely match complex textual prompts. The model stands out for its mixture-of-experts (MoEs) layers, incorporating both space-MoE and time-MoE layers, allowing for billions of unique diffusion paths. This distinct approach enables each path to function as a "painter," translating individual parts of the text into corresponding image segments with high fidelity. RAPHAEL has outperformed other state-of-the-art models like Stable Diffusion and DALL-E 2. It excels in generating images across diverse styles, such as Japanese comics and cyberpunk, and has achieved impressively low zero-shot FID scores on the COCO dataset. Training on a combination of a subset of LAION-5B and some internal datasets has ensured a broad and diverse range of training images and text for RAPHAEL.

Several tools based on diffusion models have also emerged, such as the following:

- **Imagen2** [50]: this model can generate realistic images by improving the way it pairs images with captions in its training data. The model is adept at understanding context and can edit images, including inpainting and outpainting. It also offers style conditioning, allowing for the use of reference images to guide style adherence, providing greater flexibility and control. However, it struggles with complex object placement and specific detail generation, and there is a possibility of biased content, so safety measures are essential. Trained on a large and diverse dataset, Imagen2 achieves high-quality, contextually aligned image generation.
- **Dall-E3** [51]: has made significant strides in text-to-image generation through the use of improved image captions to enhance prompt following. By developing a custom image captioner to generate detailed, synthetic captions, the model has greatly improved its ability to follow prompts, coherence, and the overall aesthetics of the generated images. However, DALL-E 3 still grapples with issues such as spatial awareness, object placement, unreliable text rendering, and the tendency to hallucinate specific details like plant species or bird types. The model's training consists of a mix of 95% synthetic captions and 5% ground truth captions, which helps regulate inputs and prevent overfitting. This thorough training process allows DALL-E 3 to produce high-quality images with improved prompt following and coherence.

As we have seen in this section, we have analysed the different approaches that are currently being researched within the domain of text-to-image synthesis. The most commonly used techniques have been GANs, Transformers, Diffusion Models and the CLIP model. This shows that there are a large number of synthetic image generation techniques that will allow the creation of large datasets created with many different techniques. This will allow the creation of detection models that are able to generalise better to real situations.

2. Image-to-Image Translation

Recent advances in image-to-image translation have introduced several cutting-edge models that enhance generated images' quality, efficiency, and versatility.

Computer vision is one of the most important fields where GANs are applied, and realistic image generation is the most widely used application of these techniques. For example, Augmented CycleGAN [52] builds on the traditional CycleGAN architecture to handle more complex image-to-image translation tasks, improving domain adaptation, style transfer, and reducing artifacts. DualGAN++ [53] introduces advanced regularization techniques and optimized training strategies, resulting in higher fidelity and fewer distortions in synthetic images. CUT++ [54] refines the original CUT model with contrastive learning techniques and enhanced loss functions for generating higher-quality synthetic images, especially in scenarios with limited data availability. SPADE++ [55] incorporates new strategies for better handling spatial inconsistencies and enhancing the realism of high-resolution synthetic images, particularly effective for images with complex structures. SSIT-GAN [56] leverages self-supervised learning techniques to generate high-quality synthetic images with self-supervised loss functions, useful for applications with limited annotated data. UMGAN [57] proposes a unified approach for multimodal image-to-image translation, enabling the generation of diverse synthetic images from multiple input modalities across various applications. Zero-shot GANs [58] aim to generate images without extensive labelled data, enhancing the zero-shot learning capabilities of GANs. This approach allows for the creation of diverse and high-quality images even with minimal training data.

Recent advancements in GAN-based synthetic image generation have been focused on enhancing image quality, efficiency, and usability across different domains. StyleGAN3 tackles the issue of "texture sticking" in generated images by introducing architectural revisions to eliminate aliasing, ensuring that image details move naturally with depicted objects. The new design interprets all signals continuously, achieving full equivariance to translation and rotation at subpixel scales. This results in images that maintain the high quality of StyleGAN2 but with improved internal representations, making StyleGAN3 more suitable for video and animation generation. The model was trained using high-quality datasets such as FFHQ, METFACES, AFHQ, and a newly collected BEACHES dataset. However, the architecture assumes specific characteristics of the training data, which can lead to challenges when these assumptions are not met, such as with aliased or low-quality images. Additionally, further improvements might be possible by making the discriminator equivariant and finding ways to reintroduce noise inputs without compromising equivariance [59], [60]. EfficientGAN [61] focuses on optimizing computational efficiency while maintaining high-quality image generation. This model aims to reduce the resource requirements for training GANs without compromising the generated images' visual quality. It introduces novel architectural modifications and training strategies that balance performance and efficiency.

Other authors have explored how to combine GANs with other types of techniques such as Latent Diffusion Models, which combine GANs with diffusion models to achieve high-resolution image synthesis. The integration of latent diffusion models helps in generating detailed and high-quality images while maintaining the robustness of GANs [62]. In contrast, Torbunov *et al.* [63] chose to combine them with Transformers. They introduced UVCGAN, an advanced model designed for image-to-image translation, focusing on synthetic image generation. This model improves upon the traditional CycleGAN framework by integrating a Vision Transformer (ViT) into the generator, enhancing its ability to learn non-local patterns. UVCGAN is highly effective for unpaired image-to-image translation tasks, making it a valuable tool for applications in fields such as art, design, and scientific simulations. ViT enables more complex and nuanced image transformations, pushing the boundaries of synthetic image generation possibilities.

Recently, significant developments have been made in Variational Autoencoders (VAEs) for synthetic image generation. These advancements have resulted in the creation of innovative models that enhance the quality, efficiency, and versatility of the images generated. For instance, Conditional VAEs [64] have improved inpainting results and training efficiency by utilizing pre-trained weights and datasets such as CIFAR-10, ImageNet, and FFHQ. VQ-VAE-2 employs hierarchical latent representations to capture high-resolution details, leading to a notable improvement in image fidelity and diversity [65]. NVAE [66], with its hierarchical architecture and advanced regularization techniques, has enabled high-resolution, realistic image generation. Another example is StyleVAE [67], which integrates VAEs with style transfer techniques to produce visually appealing images with stylistic consistency. Additionally, FHVAE has enhanced the disentanglement of latent factors, allowing for better control over image attributes [68]. EndoVAE [69], developed by Diamantis et al., introduces a fresh approach for producing synthetic endoscopic images using a Variational Autoencoder (VAE). This novel technique addresses the drawbacks of traditional GAN-based models, particularly in the domain of medical imaging where maintaining data privacy and diversity is crucial. EndoVAE is specifically designed to generate a diverse set of high-quality synthetic images, which can be used in lieu of real endoscopic images. This aids in the training of machine learning models for medical diagnosis. The outcomes illustrate that EndoVAE adeptly creates realistic endoscopic images, positioning it as a promising tool for advancing medical image analysis and circumventing the challenges stemming from limited data availability.

Furthermore, Dos Santos *et al.* [70] have introduced a Synthetic Data Generation System (SDGS) that utilizes Variational Autoencoders (VAEs) to produce synthetic images. Their system aims to automate the creation of synthetic datasets by using the Linked Data (LD) paradigm to collect and merge data from multiple repositories. The SDGS framework incorporates advanced feature engineering methods to enhance the quality of the dataset before training the VAE model. This results in synthetic images that closely mimic real-world data, making them extremely useful for training machine learning models, especially in scenarios where actual data is scarce. The system's efficacy has been confirmed through various case studies, demonstrating that the generated synthetic data achieves high accuracy and closely resembles the original datasets in crucial characteristics. Seunghwan *et al.* [71] have introduced a new method for creating synthetic data using Variational Autoencoders (VAEs). Their approach overcomes the limitations of the typical Gaussian assumption in VAEs by incorporating an infinite mixture of asymmetric Laplace distributions in the decoder. This advancement provides more flexibility in capturing the underlying data distribution, which is crucial for generating high-quality synthetic data. Their model, known as "DistVAE," has demonstrated exceptional performance in generating synthetic datasets that maintain statistical similarity to the original data and also ensures privacy preservation. The effectiveness of the approach was confirmed through experiments on various real-world tabular datasets, indicating that DistVAE can generate accurate synthetic data while allowing for adjustable privacy levels through a tunable parameter. This makes it particularly valuable in situations where data privacy is a concern.

Finally, we can see how the use of diffusion models in image-to-image translation is also beginning to be explored. For example, Parmar *et al.* [72] proposed pix2pix-zero, a method for image-to-image

TABLE III. Overview of Techniques for Detecting AI-Generated Images

Authors	Year	Technique	Target Outcome	Data Used	Open Source
Shiohara *et al.* [19]	2022	Self-blended images	Detect fake or synthetic images	Self-blended image data	Yes
Wang *et al.* [79]	2023	Diffusion Reconstruction Error	Detect difusión model-generated images	DiffusionForensics dataset	Yes
Ma *et al.* [80]	2023	Deterministic reverse and denoising computation errors	Detect images from difusión models	CIFAR-IO, TinyImageNet, CelebA	Yes
Zhong *et al.* [78]	2023	Texture patch analysis	Identify Al-generated images	Datasets from 17 generative models	Yes
lorenz *et al.* [82]	2023	Intrinsic Dimensionality-based	Detect artificial images from deep diffusion models	CiFake, ArtiFact, DiffusionDB, LAION-5B, SAC	Yes
Alzantot *et al.* [77]	2023	Wavelet-packet representation analysis	Differentiate real and synthetic images	FFHQ, CelebA, LSUN, Face Forensics++	Yes
Poredi *et al.* [75]	2023	Frequency analysis	Identify Al-generated images on social media	Stanford image dataset	Yes
Bammey *et al.* [76]	2023	Frequency artifacts analysis	Detect images generated by diffusion models	Raise and Dresden datasets	Yes
Guarnera *et al.* [83]	2023	Hierarchical classification	Identify deepfake images	CelebA, FFHQ, ImageNet	Yes
Ojha *et al.* [85]	2023	Universal fake image detector	Enhance detection of synthetic or fake images	Images generated by various models	Yes
Mathys *et al.* [86]	2024	CNN-based pixel-level analysis	Identify synthetic images	Diverse dataset with real and synthetic images	No
Coccomini *et al.* [84]	2024	Visual and textual feature classification	Detect synthetic images from diffusion models	MSCOCO and Wikimedia datasets	Yes
Tan *et al.* [87]	2024	Category Common Prompt in CLIP	Enhance detection of deepfakes	Images generated by various models	Yes
Sinitsa *et al.* [74]	2024	Fingerprint-based	Detect synthetic images with low-budget models	Various models datasets	Yes
Keita *et al.* [88]	2024	Vision-language model with dual LORA mechanism	Detect synthetic images using vision-language model	Various datasets	Yes

translation without relying on text prompts or additional training. This approach utilizes cross-attention guidance to maintain image structure and automatically discovers editing directions in the text embedding space. The architecture leverages pre-trained Stable Diffusion models for tasks like object type changes and style transformations. The model's performance is assessed using real and synthetic images from the LAION 5B dataset. However, some limitations include the low resolution of the cross-attention map for fine details and challenges with atypical poses and fine-grained edits.

In this section we have analysed the latest work in the field of Image-to-Image translation, focusing on image alterations while maintaining some visual features. Within this domain we have looked at three main approaches: GANs, AutoEncoders and diffusion models. We can observe that this domain although it has been widely explored, still presents a wide range of possibilities.

B. Detection of AI-Generated Images

The development of generative models requires the creation of detection models to differentiate between AI-generated and real images. Detection methods can be split into two main types: those focused solely on improving detection performance and those that enhance detectors with additional features such as generalizability, robustness, and interpretability while maintaining accurate and effective detection capabilities. An overview of techniques for detecting AI-generated images is provided in Table III, summarizing various methods and their key features, including the application areas and datasets used. For example, the Deep Image Fingerprint (DIF) [74] method is specifically designed to detect low-budget synthetic images. It can identify images generated by both Generative Adversarial Networks (GANs) and Latent Text-to-Image Models (LTIMs). The method utilizes datasets from various models, including CycleGAN, ProGAN, BigGAN, StyleGAN, Stable Diffusion, DALL·E-2, and GLIDE, and achieves high detection accuracy with minimal training samples. While it excels in detecting synthetic images, it may encounter some challenges with models like GLIDE and DALL·E-2 due to their weaker, less distinct fingerprints.

Some authors still opt for more traditional techniques, such as the **Fourier Transform** for the detection of artefacts left in the image samples. For example, the AUSOME (AUthenticating SOcial MEdia) [75] method is focused on identifying AI-generated images on social media. It achieves this by utilizing frequency analysis techniques, such as the Discrete Fourier Transform (DFT) and Discrete Cosine Transform (DCT), to compare the spectral features of AI-generated images, like those produced by DALL-E 2, with legitimate images from the Stanford image dataset. AUSOME can distinguish between AI-generated and real images by examining differences in frequency responses. Although it demonstrates high accuracy, it may encounter difficulties when dealing with images where semantic content is essential for determining authenticity. Nevertheless, this method presents a promising approach for verifying social media images, particularly in light of the increasing prevalence of AI-generated content. Synthbuster [76] is a technique developed to identify images created by diffusion models by analyzing frequency artifacts in the Fourier transform of residual images. This method is effective at spotting synthetic images, even when they are slightly compressed in JPEG format, and it works well with unknown models. It analyzes real images from the RAISE and Dresden datasets and synthetic images from various models such as Stable Diffusion, Midjourney, Adobe Firefly, DALL·E 2, and DALL·E 3. While Synthbuster is generally effective, it may encounter challenges when dealing with different compression levels and diverse image categories.

Other authors focus on taking advantage of **textures**, in order to exploit all available information. For instance, Alzantot *et al.* [77] proposed multi-scale wavelet-packet representations. Their deepfake image analysis and detection technique aims to differentiate real from synthetic images by analyzing their spatial and frequency information. This method has undergone evaluation using various datasets, including FFHQ, CelebA, LSUN, and FaceForensics++. It has shown strong capabilities in identifying GAN-generated images, such as those created by StyleGAN. However, it may face challenges when analyzing complex images where semantic information is crucial, and its effectiveness may be limited to the detection of image-based synthetic media. PatchCraft [78] introduces a fresh approach to identifying synthetic AI-generated images. Instead of relying solely on global semantic information, this method focuses on analyzing texture patches within the images for more effective detection. To enhance detection, the method employs a preprocessing step called Smash&Reconstruction, which removes global semantic details and amplifies texture patches, thereby utilizing the contrast between rich and poor texture regions to boost performance. Tested on datasets from 17 common generative models, including ProGAN, StyleGAN, BigGAN, CycleGAN, ADM, Glide, and Stable Diffusion, the method has shown superior adaptability and resilience against previously unseen models and image distortions. Nevertheless, it may encounter challenges when dealing with images in which semantic information is critical for accurate detection.

An analysis on the **error** inserted in generated images has also been a productive research line. For example, the DIRE (DIffusion REconstruction Error) [79] method is utilized to identify images created through diffusion processes by comparing the reconstruction error between an original image and its reconstructed version using a pre-trained diffusion model. This technique is based on the idea that diffusion-generated images can be accurately reconstructed using diffusion models, unlike genuine images. DIRE has been evaluated using the DiffusionForensics dataset, encompassing images from various diffusion models, including ADM, DDPM, and iDDPM. It has demonstrated notable accuracy in detecting images and is resilient to unseen diffusion models and alterations. Nonetheless, it may encounter difficulties with the intricate features of real images. Shiohara *et al.* [19] has introduced an innovative approach for detecting fake or synthetic images, specifically deepfakes. They utilize self-blended images (SBIs) as synthetic training data to enhance the robustness of detection models. This allows the models to effectively identify various types of deepfake manipulations by scrutinizing inconsistencies and artifacts in the images. Consequently, this method provides a robust tool for preserving the authenticity of digital media in the face of increasingly advanced generative techniques. The SeDID [80] method utilizes deterministic reverse and denoising computation errors found in diffusion models. This approach includes two branches: the statistical-based SeDIDStat and the neural network-based SeDIDNNs. SeDID was evaluated on various datasets like CIFAR-10, TinyImageNet, and CelebA and demonstrated superior detection accuracy and robustness against unseen diffusion models and perturbations. However, the method may encounter challenges when dealing with the complex features of real images. Nevertheless, SeDID underscores the importance of selecting the optimal timestep to enhance detection performance.

As expected, another approach widely used by state-of-the-art researchers is **Convolutional Neural Networks**, which have demonstrated excellent performance on numerous similar classification problems [81], making it one of the most explored techniques. Some authors continue to rely on classical architectures such as ResNet. It continues to perform competitively on many classification problems. Among them, The multi-local Intrinsic Dimensionality (multiLID) [82] method is developed to identify artificial images produced by deep diffusion models. This method utilizes the local intrinsic dimensionality of feature maps extracted by an untrained ResNet18, making it efficient and not relying on pre-trained models. It has been evaluated on various datasets like CiFake, ArtiFact, DiffusionDB, LAION-5B, and SAC, demonstrating high accuracy in detecting artificial images from models including Glide, DDPM, Latent Diffusion, Palette, and Stable Diffusion. However, multiLID may have limitations in its ability to perform well on unfamiliar data from different datasets or models within the same domain. Guarnera *et al.* [83] developed a hierarchical multi-level approach for detection and identification of deepfake images produced by GANs and Diffusion Models (DMs). This method utilizes ResNet-34 models at three levels of classification: distinguishing genuine images from AI-generated ones, discerning between GANs and DMs, and identifying specific AI architectures. Their dataset comprises authentic images from CelebA, FFHQ, and ImageNet, as well as synthetic images from nine GAN models (e.g., AttGAN, CycleGAN, ProGAN, StyleGAN, StyleGAN2) and four diffusion models (e.g., DALL-E 2, GLIDE, Latent Diffusion), totalling 42,500 synthetic and 40,500 real images. With an accuracy of over 97%, the method demonstrates strong performance, but it may encounter challenges related to real-world robustness, such as JPEG compression and complex image features.

However, other authors have opted for different architectures rather than CNNs. Coccomini *et al.* [84] investigate the detection of synthetic images generated by diffusion models, such as those created with Stable Diffusion and GLIDE. Their approach involves using classifiers like multi-layer perceptrons (MLPs) and convolutional neural networks (CNNs) to distinguish synthetic images from real ones. The model is trained on datasets like MSCOCO and Wikimedia, focusing on leveraging visual and textual features for effective detection. A notable limitation of the study is the challenge of cross-method generalization, where models trained on one type of synthetic image struggle to detect images generated by different methods. This work underscores the complexities of detecting AI-generated images, particularly as diffusion models become more sophisticated. Ojha et al. [85] have introduced a method to enhance the detection of synthetic or fake images generated by various models, including **GANs** and diffusion models. Their approach aims to create a universal fake image detector that performs well across different generative models. This is achieved through a combination of convolutional neural networks (CNNs) and advanced training techniques to identify subtle anomalies commonly found in AI-generated images. The model is trained on diverse datasets, incorporating images generated by various models to improve its reliability. However, the study highlights a challenge in maintaining high detection accuracy when faced with new generative models not included in the training set, indicating the need for further improvements to achieve universal detection capabilities. Mathys *et al.* [86] present a method for identifying synthetic images produced by AI models. The focus is on spotting subtle artifacts and inconsistencies that are indicative of AI-generated content. Their proposed architecture utilizes a convolutional neural network to scrutinize pixel-level details and capture the distinct markers left by generative models. Training the model on a diverse dataset containing both real and synthetic images from various sources makes it adept at generalizing across different types of AI-generated content. This method significantly boosts the accuracy of detecting fake images, effectively tackling the challenges brought about by the increasingly lifelike outputs of modern generative models. This research holds particular significance in upholding the authenticity and integrity of digital content in an age where synthetic media is increasingly prevalent.

Lastly, we will analyse some research that has chosen other novel approaches such as the use of models like **CLIP** or **vision-language**

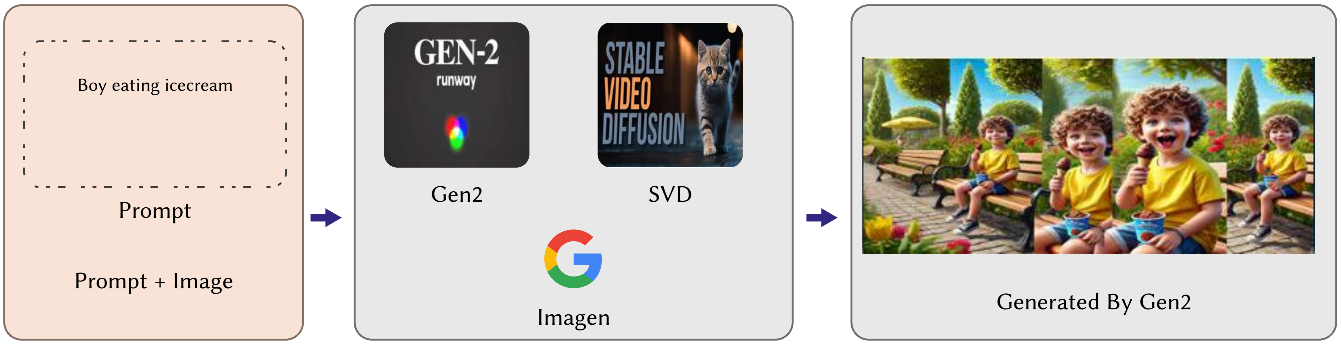


Fig. 5. Overviewof the main approaches to video generation with AI.

models. Tan *et al.* [87] introduce C2P-CLIP, a novel approach designed to enhance the detection of AI-generated images, specifically deepfakes, by injecting a Category Common Prompt (C2P) into the CLIP model. CLIP (Contrastive Language-Image Pre-training) is a powerful model trained on various image-text pairs, which allows it to understand and match images and text descriptions effectively. However, its application to deepfake detection has been limited by its generalization capability across different types of manipulations. The C2P-CLIP method addresses this limitation by incorporating a category-specific prompt that captures standard features across related deepfakes, improving the model's ability to generalize beyond the specific types of manipulations seen during training. This technique leverages the extensive pre-training of CLIP while fine-tuning its capacity to identify subtle inconsistencies and artifacts introduced by deepfake generation techniques. Through comprehensive experiments, the authors demonstrate that C2P-CLIP significantly outperforms existing methods on several benchmark datasets, showing superior performance in detecting a wide range of AI-generated manipulations. Keita et al. [88] present Bi-LORA, a vision-language approach designed to detect synthetic images. Bi-LORA effectively captures the unique features and artefacts of AI-generated images by leveraging a dual Low-Rank Adaptation (LORA) mechanism within a vision-language model. The method integrates visual and textual information, enhancing its ability to differentiate between real and synthetic content more accurately. Through extensive experiments, Bi-LORA demonstrates significant improvements in detection performance over traditional methods, highlighting its potential as a robust tool for identifying AI-generated images across various datasets.

Lastly, we have analysed the most recent research into the detection of synthetic image. This field is highly dependent on the previous one, as quality datasets will be needed, i.e. with intra-class variability, enough quality and resolution, and representativeness, allowing the creation of models that can be used in real situations. In this domain we have seen that the main approaches explored by researchers are CNNs, and vision-language models, although other more traditional approaches are still used.

IV. Video Generation and Detection

In recent years, the field of video generation has attracted significant attention, due to advancements in artificial intelligence, machine learning, and the emergence of diffusion models (see Fig. 5), this has forced researchers to develop new techniques to detect these synthetic samples. This section provides an overview of the current state of video generation methods, which are increasingly being used to create high-quality, realistic videos across different applications. Additionally, it explores the challenges and methods associated with detecting AI-generated videos, an area of growing importance as these technologies become more sophisticated. The aim of this section is to provide a comprehensive understanding of the methods and techniques involved in future video content creation and analysis.

A. Video Generation

In video content creation, generative models are beginning to revolutionize production and consumption by automating the generation of realistic and high-quality videos. Recently, a surge of generative video models capable of various video creation tasks has emerged. In this section we are going to analyse five different approaches: *Text-to-video*, deep learning techniques that generate synthetic video samples from text descriptions; *image-to-video* techniques that transform static images to dynamic video; *video-to-video*, a set of techniques focused on the generation of realistic video sequences by transforming or translating visual information from one video domain to another; *Text-Image-to-Video* which generates synthetic video samples from a real image and a text description; *Multimodal video generation*, this field focuses not only on the generation of the visual part of the video but also on the audio part of the video, from different inputs, such as text, image, video or audio. Deep learning-based generative models such as GANs, Variational Autoencoders (VAEs), autoregressive, and diffusion-based models have remarkably succeeded in generating realistic and diverse content. By training on large datasets, these models learn the underlying data distribution, enabling them to generate samples that closely resemble the original data. Fig. 6 illustrates the various categories of video generation.

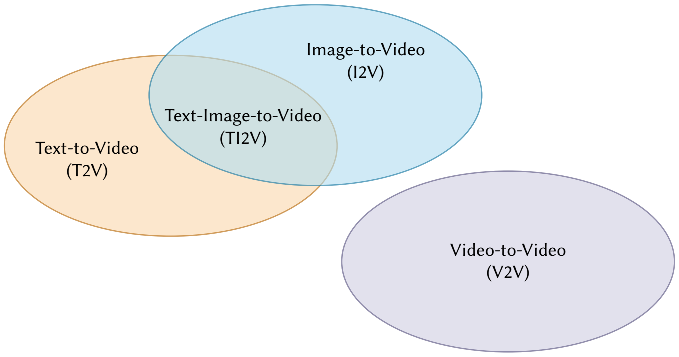


Fig. 6. Categories of video generation methods.

1. Text-to-Video Synthesis

Generating photo-realistic videos presents significant challenges, particularly when it comes to maintaining high fidelity and continuity of motion over extended sequences. Despite these difficulties, recent advancements have utilized diffusion models to enhance the realism of video generation. Text, being a highly intuitive and informative form of instruction, has become a central tool in guiding video synthesis, leading to the development of **Text-to-video (T2V)** generation models. This approach focuses on creating high-quality videos based on text descriptions, acting as a conditional input for the video generation process.

To address the challenges in text-to-video synthesis, existing methods primarily extend **Text-to-image** models by incorporating temporal modules, such as **temporal convolutions and temporal attention**, to establish temporal correlations between video frames. A notable example is the work by Ho *et al.* [89], who introduced Video Diffusion Models (VDM). This model extends text-to-image diffusion models to video generation by training jointly on both image and video data. Their approach utilizes a U-Net-based architecture, which integrates joint image-video denoising losses, ensuring temporal coherence by conditioning on both past and future frames, thus resulting in smoother transitions and more consistent motion. Building on this foundation, Ho et al. [90] proposed Imagen Video, a novel approach for generating high-definition videos using diffusion models. Imagen Video employs a cascaded video diffusion model approach, adapting techniques from text-to-image generation, such as a frozen T5 text encoder and classifier-free guidance, to the video domain. It uses a hierarchical approach, beginning with a low-resolution video to capture the overall structure and motion, which is then progressively refined to higher resolutions. Temporal dynamics are managed by conditioning each frame on previous frames, ensuring consistency throughout the video. Super-resolution techniques are subsequently applied to enhance the detail and quality of each frame.

In a different approach, Singer *et al.* [91] introduced Make-A-Video, which generates videos from textual descriptions without relying on paired text-video data. This methodology builds upon a text-to-image synthesis model and incorporates spatio-temporal layers to extend it into the video domain. The approach integrates pseudo-3D convolutional and attention layers to manage spatial and temporal dimensions efficiently. Additionally, super-resolution networks are employed to improve visual quality, and a frame interpolation network is used to increase the frame rate and smooth out the video output. Meanwhile, Zhou *et al.* [92] presented MagicVideo, a framework designed to generate high-quality video clips from textual descriptions. Instead of directly modeling the video in visual space, MagicVideo leverages a pre-trained **Variational autoencoder (VAE)** to map video clips into a low-dimensional latent space, where the distribution of videos' latent codes is learned via a diffusion model. This approach optimizes computational efficiency and improves video synthesis by performing the diffusion process in the latent space. Further pushing the boundaries of video generation, Dan Kondratyuk *et al.* [5] proposed VideoPoet, an advanced language model for zero-shot video generation. This model integrates the MAGVIT-v2 [93] tokenizer for images and videos and the SoundStream [94] tokenizer for audio, enabling the processing and generation of multimedia content within a unified framework. VideoPoet employs a prefix language model with a decoder-only architecture as its backbone, facilitating the creation of high-quality videos from textual prompts, along with interactive editing capabilities. VideoPoet is trained on a diverse set of tasks without needing paired video-text data, allowing it to learn effectively from video-only examples. It can generate videos based on textual descriptions, animate static images, apply styles [95] to videos through optical flow and depth prediction, and even extend video sequences by iteratively predicting subsequent frames.

In another innovative approach, Girdhar *et al.* [96] introduced EMU VIDEO, a **two-stages Text-to-video generation model**: first, it generates an image from text, and then it produces a video using both the text and the generated image. This method simplifies video prediction by leveraging a pretrained text-to-image model and freezing spatial layers while adding new temporal layers for video generation. EMU VIDEO efficiently achieves high-resolution video generation, maintaining the conceptual and stylistic diversity learned from large image-text datasets. Similarly, Wang *et al.* [97] proposed LaVie, a cascaded framework for Video Latent Diffusion Models (V-LDMs) conditioned on text descriptions. LaVie is composed of three networks: a base T2V model for generating short, low-resolution key frames, a Temporal interpolation (TI) model for increasing the frame rate and enriching temporal details, and a Video super-resolution model (VSR) for enhancing the visual quality and spatial resolution of the videos. The base T2V model modifies the original 2D UNet to handle spatio-temporal distributions and utilizes joint fine-tuning with both image and video data to prevent catastrophic forgetting, resulting in significant video quality improvements. The TI model uses a diffusion UNet to synthesize new frames, enhancing video smoothness and coherence, while the VSR model adapts a pre-trained image upscaler with additional temporal layers, enabling efficient training and high-quality video generation.

Further developments include the work by Menapace *et al.* [98], who proposed a method to generate high-resolution videos by modifying the **Efficient diffusion model (EDM)** [99] framework for high-dimensional inputs and developing a scalable transformer architecture inspired by Far-reaching interleaved transformerss (FITs) [100]. They adjust the EDM framework to handle high SNR in videos with a scaling factor for optimal denoising. This method addresses the scarcity of captioned video data by jointly training the model on both images and videos, allowing for more effective learning of temporal dynamics. The video generation uses FITs, transformer models that reduce complexity by compressing inputs with learnable latent tokens and employing cross-attention and self-attention to focus on spatial and temporal information. The approach includes conditioning tokens for text and metadata and uses a cascade model: the first stage generates low-resolution videos, and the second stage refines them into high-resolution outputs. During training, variable noise levels are introduced to the second-stage inputs to improve upsampling quality, aiming for effective high-quality video generation. In addressing data scarcity, Chen *et al.* [101] designed VideoCrafter2, a model that improves spatio-temporal consistency in video diffusion models through a data-level disentanglement strategy. This approach separates motion aspects from appearance features, leveraging low-quality videos for motion learning and high-quality images for appearance learning. This design strategy eases a targeted fine-tuning process with high quality images, with the aim of significantly increasing the visual fidelity of the generated content without compromising the precision of motion dynamics. Importantly, synthetic images with complex concepts are used for finetuning, rather than real images, to enhance the concept composition ability of video models.

Furthermore, Ma *et al.* [102] introduced Latte, a simple and general video diffusion method that extends **Latent diffusion models (LDMs)** for video generation by employing a series of transformer blocks to process latent space representations of video data obtained from a pre-trained variational autoencoder. Latte specifically addresses the inherent disparities between spatial and temporal information in videos by decomposing these dimensions, allowing for more efficient processing. The method includes four efficient Transformer-based model variants, designed to manage the large number of tokens extracted from input videos, thereby improving the overall performance and scalability of video generation. Li *et al.* [103] introduced VideoGen, a text-to-video generation method that produces high-definition videos with strong frame fidelity and temporal consistency using reference-guided latent diffusion. In their approach, an off-the-shelf T2I model like Stable diffusion (SD) generates a high-quality image from a text prompt, which then serves as a reference for video generation. This process involves a cascaded latent diffusion module conditioned on both the reference image and text prompt, followed by a flow-based temporal upsampling step that enhances temporal resolution. Finally, a video decoder maps the latent video representations into high-definition videos, improving visual fidelity and reducing artifacts while focusing on learning video dynamics. The training process benefits from high-quality unlabeled video data, using the first frame of a ground-truth video as the reference image to enhance motion smoothness and realism.

Building on the VQ-VAE architecture, Godiva *et al.* [104] proposed GODIVA, an open-domain text-to-video model pre-trained on the HowTo100M [105] dataset. This model generates videos in an auto-regressive manner using a three-dimensional sparse attention mechanism. Initially, a VQ-VAE auto-encoder represents continuous video pixels as discrete video tokens. Subsequently, the three-dimensional sparse attention model utilizes language input alongside these discrete video tokens to generate videos, effectively considering temporal, column, and row information. Similarly, Ding *et al.* [106] advanced the field by introducing CogVideo, a 9B-parameter transformer built upon the pretrained text-to-image model CogView2 [42] for video generation. CogVideo employs a multi-frame-rate hierarchical training strategy, which aligns text with video clips by controlling frame generation intensity and ensuring accurate alignment between text and video content. This is achieved by prepending text prompts with frame rate descriptions, which significantly enhances generation accuracy, particularly for complex semantic movements. Additionally, CogVideo's dual-channel attention mechanism improves the coherence of generated videos by focusing on both textual and visual cues simultaneously. This approach allows CogVideo to efficiently adapt a pretrained model for video synthesis without the need for costly full retraining.

Expanding on the capabilities of earlier models, Wu *et al.* [107] developed NUWA, a unified **multimodal pre-trained model** designed for generating and manipulating visual data, including images and videos, across various visual synthesis tasks. NUWA utilizes a 3D transformer **encoder-decoder** framework to process 1D text, 2D images, and 3D videos. This model introduces a 3D nearby attention (3DNA) mechanism that efficiently handles visual data, reduces computational complexity, and enables high-quality synthesis with notable zero-shot capabilities. Further advancing this work, Wu *et al.* [108] introduced NUWA-Infinity, a groundbreaking model for infinite visual synthesis capable of generating high-resolution images or long-duration videos of arbitrary size. The model features an autoregressive over autoregressive generation mechanism, with a global patch-level model managing inter-patch dependencies and a local token-level model handling intra-patch dependencies. To optimize efficiency, NUWA-Infinity incorpores a Nearby context pool (NCP) to reuse previously generated patches, minimizing computational costs while maintaining robust dependency modeling. Additionally, an Arbitrary direction controller (ADC) enhances flexibility by determining optimal generation orders and learning position embeddings tailored for diverse synthesis tasks. NUWA-Infinity thus transcends the limitations of fixed-size approaches, enabling comprehensive and efficient content creation on a variable scale. In contrast to these approaches, Yan *et al.* [109] proposed VideoGPT, a simpler and more efficient architecture for scaling likelihood-based generative modeling to natural videos. By employing VQ-VAE with 3D convolutions and axial self-attention, VideoGPT learns downsampled discrete latent representations of raw videos. These representations are then autoregressively modeled by a GPT-like architecture with spatio-temporal position encodings to generate videos. This method involves training a VQ-VAE with an encoder that downsamples space-time and a decoder that upsamples it, sharing spatio-temporal embeddings across attention layers. Furthermore, a prior over the VQ-VAE latent codes is learned using an Image-GPT-like architecture with dropout for regularization, which enables conditional sample generation via cross attention and conditional norms. Blattmann *et al.* [110] introduced a novel approach to efficient high-resolution video generation through Video LDMs, by adapting pre-trained image diffusion models into video generators. They achieve this by temporal fine-tuning with alignment layers, which maintains computational efficiency. Initially, an LDM is pre-trained on images and then transformed into a video generator by adding a temporal dimension and fine-tuning on video sequences. Additionally, diffusion model upsamplers are temporally aligned for consistent video super resolution, allowing the efficient training of high-resolution, long-term consistent video generation models using pre-trained image LDMs with added temporal alignment.

Building on these advancements, Chen *et al.* [111] introduced two diffusion models for high-quality video generation: T2V and Image-to-video (I2V). The T2V model, based on SD 2.1, incorporates temporal attention layers to ensure temporal consistency and employs a joint image and video training strategy. The VideoCrafter T2V model further leverages a Latent Video Diffusion Model (LVDM) with a video VAE and a video latent diffusion model, where the VAE reduces sample dimensions to improve efficiency. Video data is encoded into a compressed latent representation, processed through a diffusion model with noise added at each timestep, before being decoded by the VAE to generate the final video. He *et al.* [112] expanded on the concept of video generation by introducing a hierarchical LVDM framework that extends videos beyond the training length. Their method addresses performance degradation with conditional latent perturbation and unconditional guidance. Their lightweight video diffusion models use a low-dimensional 3D latent space, significantly outperforming pixel-space models with limited computational resources. By compressing videos into latents using a video autoencoder and utilizing a unified video diffusion model for both unconditional and conditional generation, their approach generates videos autoregressively and improves coherence and quality over extended lengths with hierarchical diffusion.

To further advance video generation, Wang *et al.* [113] proposed **ModelScope Text-to-Video (ModelScopeT2V)**, a simple yet effective baseline for video generation. This model introduces two key technical contributions: a spatio-temporal block to model temporal dependencies in text-to-video generation, and a multi-frame training strategy with both image-text and video-text paired datasets to enhance semantic richness. ModelScopeT2V evolves from a text-to-image model (stable diffusion) and includes spatio-temporal blocks to ensure consistent frame generation and smooth transitions, adapting to varying frame numbers during training and inference. In the realm of scalable and efficient video generation, Gupta *et al.* [114] proposed W.A.L.T, a simple yet scalable and efficient transformer-based framework for latent video diffusion models. Their approach consists of two stages: an autoencoder compresses images and videos into a lower-dimensional latent space, allowing for efficient joint training on combined datasets. Subsequently, the transformer employs window-restricted self-attention layers that alternate between spatial and spatio-temporal attention, reducing computational demands and supporting joint image-video processing. This method facilitates high-resolution, temporally consistent video generation from textual descriptions, offering an innovative approach to T2V synthesis. Villegas et al. [115] contributed to the field by proposing Phenaki, a unique C-ViViT encoder-decoder structure for generating variable-length videos from textual inputs. This model compresses video data into compact tokens, allowing for the production of coherent and detailed videos. By utilizing a bidirectional masked transformer to translate text tokens into video tokens, the model can generate long, temporally coherent videos from both open-domain and sequential prompts. It

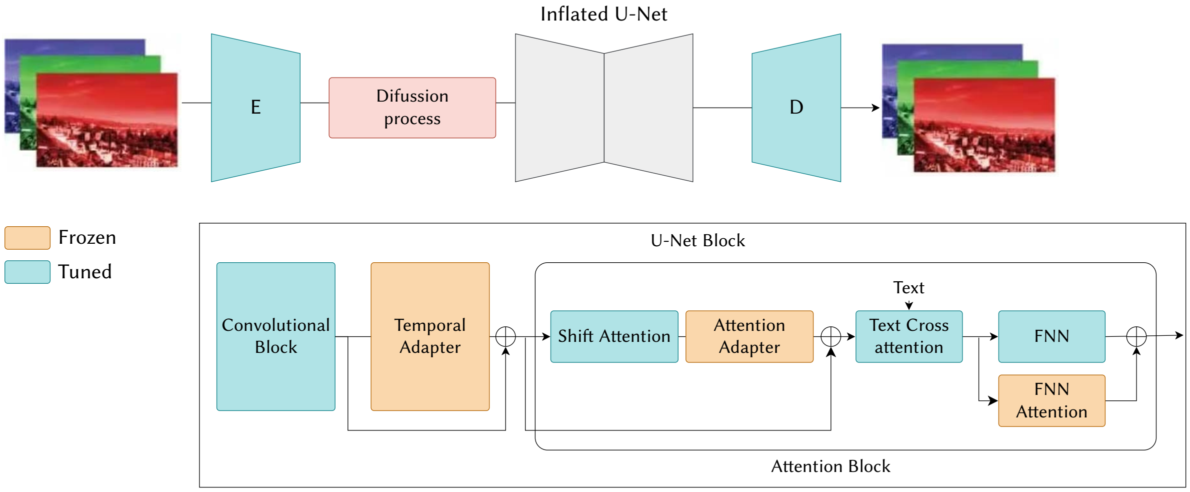


Fig. 7. SimDA [116] architecture.

also improves video token compression by 40% by exploiting temporal redundancy, enhancing reconstruction quality and accommodating variable video lengths, while the causal variation of ViViT manages temporal and spatial dimensions in an auto-regressive manner.

Previous methods of text-to-video generation face high computational costs with pixel-based VDMs or struggle with text-video alignment with latent-based VDMs. To marry the strength and alleviate the weakness of pixel-based and latent-based VDMs, Zhang *et al.* [117] proposed Show-1, a hybrid model that combines both pixel-based and latent-based VDMs to overcome the limitations of previous methods. By employing pixel-based VDMs to create low-resolution videos with strong text-video correlation, and then using latent-based VDMs to upsample these to high resolution, Show-1 ensures precise text-video alignment, natural motion, and high visual quality with reduced computational cost. Khachatryan *et al.* [118] built upon the Stable diffusion T2I model to develop Text2Video-Zero, a zero-shot T2V synthesis model. This approach enriches latent codes with motion dynamics to ensure temporal consistency and employs a cross-frame attention mechanism to maintain object appearance and identity across frames. Although Text2Video-Zero enables high-quality, temporally consistent video generation from textual descriptions without additional training, leveraging existing pre-trained T2I models, there is still potential for improvement. It struggles to generate longer videos with sequences of actions.

Furthermore, FuWeng *et al.* [119] introduced ART•V, an efficient framework for **autoregressive video generation** using diffusion models. ART•V generates frames sequentially, conditioned on previous frames, by focusing on simple, continuous motions between adjacent frames, which helps to avoid the complexity of long-range motion modeling. This approach retains the high-fidelity generation capabilities of pre-trained image diffusion models with minimal modifications and can produce long videos from diverse prompts, such as text and images. To address the common issue of drifting in autoregressive models, ART•V incorporates a masked diffusion model that draws information from reference images rather than relying solely on network predictions, thereby reducing inconsistencies. By conditioning on the initial frame, ART•V enhances global coherence, which is particularly useful for generating long videos. The framework also employs a T2I-Adapter for conditional generation, ensuring high fidelity with minimal changes to the pre-trained model, matching the inference speed of one-shot models, and supporting larger batch sizes during training. In summary, ART•V effectively reduces drifting issues in video generation by incorporating masked diffusion, anchored conditioning, and noise augmentation to better align training with testing. Shi *et al.* [120] introduced BIVDiff, a training-free video synthesis framework that integrates frame-wise video generation, mixed inversion, and temporal smoothing. This framework bridges the gap between specific image diffusion models (e.g., ControlNet, Instruct Pix2Pix) and general text-to-video diffusion models (e.g., VidRD, ZeroScope). The process begins with frame generation using an image diffusion model, followed by Mixed Inversion to adjust latent distributions, which balances temporal consistency with the open-generation capability of video diffusion models. Finally, video diffusion models are applied for temporal smoothing. This method effectively addresses issues of temporal consistency and task generalization that are common in previous training-free approaches.

Finally, Xing *et al.* [116] proposed a parameter-efficient video diffusion model called Simple Diffusion Adapter (SimDA), see Fig. 7, which fine-tunes the large T2I model (i.e., Stable Diffusion) for enhanced video generation. SimDA generates videos from textual prompts through efficient one-shot fine-tuning of pre-trained Stable Diffusion models, focusing on a parameter-efficient approach by fine-tuning only 24 million out of the 1.1 billion parameters. The model employs an adapter with two learnable fully connected layers, incorporating spatial adapters to capture appearance transferability and temporal adapters to model temporal information, utilizing GELU activations and depth-wise 3D convolutions. Additionally, SimDA introduces Latent-shift attention (LSA) to replace the original spatial attention, enhancing temporal consistency without adding new parameters. More recently, Qing *et al.* [121] presented HiGen, a diffusion-based model that improves video generation by decoupling spatial and temporal factors at both the structure and content levels. At the structural level, HiGen splits the T2V task into spatial reasoning, which involves generating spatially coherent priors from text, and temporal reasoning, which creates temporally coherent motions from these priors using a unified denoiser. On the content side, HiGen extracts cues for motion and appearance changes from input videos to guide training, thereby enhancing temporal stability and allowing for flexible content variations. Despite its strengths, HiGen faces challenges in generating detailed objects and accurately modeling complex actions due to computational and data quality limitations.

As we have seen in this section, for the generation of video from text, the main approaches used are the application of T2I techniques together with temporal modules, attention mechanisms, transformers and autoencoder. However, in recent years many researchers are focusing on diffusion models, which are becoming more and more widely used and are expected to increase in popularity in the coming years.

2. Image-to-Video Synthesis

Generating videos from static images poses significant challenges, particularly in preserving temporal consistency and achieving realistic motion across frames. Despite these difficulties, advancements in image-to-video synthesis have leveraged sophisticated modeling techniques to transform still images into dynamic video sequences. This area has become increasingly important for various applications, ranging from content creation to enhanced video editing tools.

Recent methods in image-to-video synthesis focus on generating high-quality videos by incorporating temporal dynamics into the transformation process. Techniques like temporal modeling and attention mechanisms are employed to ensure smooth transitions between frames, thus maintaining coherence and realism in the generated videos. A noteworthy contribution to this field is the work by Wu *et al.* [122], which introduces LAMP, a few-shot-based tuning framework for Text-to-video generation, leveraging a first-frame-attention mechanism to transfer information from the initial frame to subsequent ones. This approach, which focuses on fixed motion patterns, is constrained in its ability to generalize across diverse scenarios. LAMP utilizes an off-the-shelf text-to-image model for content generation while emphasizing motion learning through expanded pre-trained 2D convolution layers and modified attention blocks for temporal-spatial motion learning. A first-frame-conditioned pipeline ensures high video quality by retaining the initial frame's content and applying noise to subsequent frames during training. During inference, high-quality first frames generated by SD-XL enhance video performance. Despite its promise, LAMP faces challenges with complex motions and background stability, suggesting areas for future improvement. Guo *et al.* [123] introduced the I2V-Adapter, a lightweight and plug-and-play solution designed for text-guided Image-to-video generation. The key innovation of this adapter lies in its cross-frame attention mechanism, which preserves the identity of the input image by propagating the unnoised image to subsequent noised frames. This approach ensures compatibility with pretrained Text-to-video models, maintaining their weights unchanged while seamlessly integrating the adapter. By introducing minimal trainable parameters, the I2V-Adapter not only reduces training costs but also ensures smooth compatibility with community-driven models and tools. Moreover, the authors incorporated a Frame Similarity Prior, which provides adjustable control coefficients to balance motion amplitude and video stability, thereby enhancing both the controllability and diversity of the generated videos.

Futhermore, Zhang *et al.* [124] proposed MoonShot, a video generation model that leverages both image and text as conditional inputs. MoonShot addresses limitations in controlling visual appearance and geometry by employing the Multimodal video block (MVB) as its core component. This module integrates spatial-temporal layers for comprehensive video feature representation and utilizes a decoupled cross-attention layer to condition both image and text inputs effectively. Notably, MoonShot reuses pre-trained weights from text-to-image models, allowing for the integration of pre-trained image ControlNet modules to achieve geometry control without necessitating additional training. The model's architecture, which includes spatial-temporal U-Net layers and decoupled multimodal cross-attention layers, ensures high-quality frame generation and temporal consistency. As a result, MoonShot is versatile, supporting tasks like image animation and video editing without the need for fine-tuning, while also enabling geometry-controlled generation through the effective integration of ControlNet modules. Gong *et al.* [125] proposed AtomoVideo, a high-fidelity Image-to-video generation framework that transforms product images into engaging promotional videos. AtomoVideo achieves superior motion intensity and consistency compared to existing methods and can also perform Text-to-video generation by combining advanced text-to-image models. The approach involves using a pre-trained T2I model with added temporal convolution and attention modules, training only the temporal layers, and injecting image information at two positions: low-level details via VAE encoding and high-level semantics via CLIP image encoding and cross-attention. Long video frames are predicted iteratively, using initial frames to generate subsequent ones. The framework is trained using Stable Diffusion 1.5 and a 15M internal dataset, employing zero terminal SNR and v-prediction techniques for stability. During inference, classifier-free guidance with image and text prompts significantly enhances the stability of the generated output.

Other researchers have explored diffusion models for the creation of videos from images. For example, Shi *et al.* [126] proposed Motion-I2V, a novel framework for consistent and controllable text-guided image-to-video generation. Unlike previous methods, Motion-I2V factorizes the process into two stages with explicit motion modeling. The first stage involves a diffusion-based motion field predictor to deduce pixel trajectories of the reference image. The second stage introduces motion-augmented temporal attention to enhance the limited 1-D temporal attention in video latent diffusion models, effectively propagating reference image features to synthesized frames guided by predicted trajectories. By training a sparse trajectory ControlNet for the first stage, Motion-I2V enables precise control over motion trajectories and regions, also supporting zero-shot Video-to-video translation. Although Motion-I2V provides fine-grained control of I2V generation through sparse trajectory guidance, region-specific animation and zero-shot Video-to-video translation, it is limited in handling occlusions, brightness uniformity and complex motion.

Expanding on the idea of temporal consistency, Ren *et al.* [127] proposed ConsistI2V, a diffusion-based method for I2V generation, designed to enhance visual consistency by using spatiotemporal attention over the first frame to maintain spatial and motion coherence. They introduced FrameInit, an inference-time noise initialization strategy that uses the low-frequency band from the first frame to stabilize video generation, which supports applications such as long video generation and camera motion control. The approach leverages cross-frame attention mechanisms and local window temporal layers to achieve fine-grained spatial conditioning and temporal smoothness. The ConsistI2V's architecture, based on a U-Net structure adapted with temporal layers, employs a latent diffusion model to generate videos that closely align with the first frame and follow the textual description. To address motion consistency and efficiency, Shen *et al.* [128] proposed a novel approach to Conditional image-to-video (cI2V) generation by disentangling RGB pixels into spatial content and temporal motions. Using a 3D-UNet diffusion model, they predict temporal motions, including motion vectors and residuals, to improve consistency and efficiency. The approach begins with Decouple-Based Video Generation (D-VDM) to predict differences between consecutive frames and is further refined with Efficient Decouple-Based Video Generation (ED-VDM), which separates content and temporal information using motion vectors and residuals extracted via CodeC. The model employs Gaussian noise and a diffusion model to learn the video distribution score and generate a video clip from the initial frame and text condition. The approach includes Decoupled Video Diffusion Model using DDPM to estimate video distribution scores and a ResNet bottleneck module to encode the first frame, improving spatial and temporal representation alignment. Efficient representation is achieved using I-frames and P-frames, with compression via a Latent Diffusion autoencoder, optimizing video generation through a learned joint distribution of motion vectors and residuals.

Facing the challenge of maintaining temporal coherence while preserving detailed information about the characters in the image-video synthesis for character animation is difficult. Hu *et al.* [129] proposed

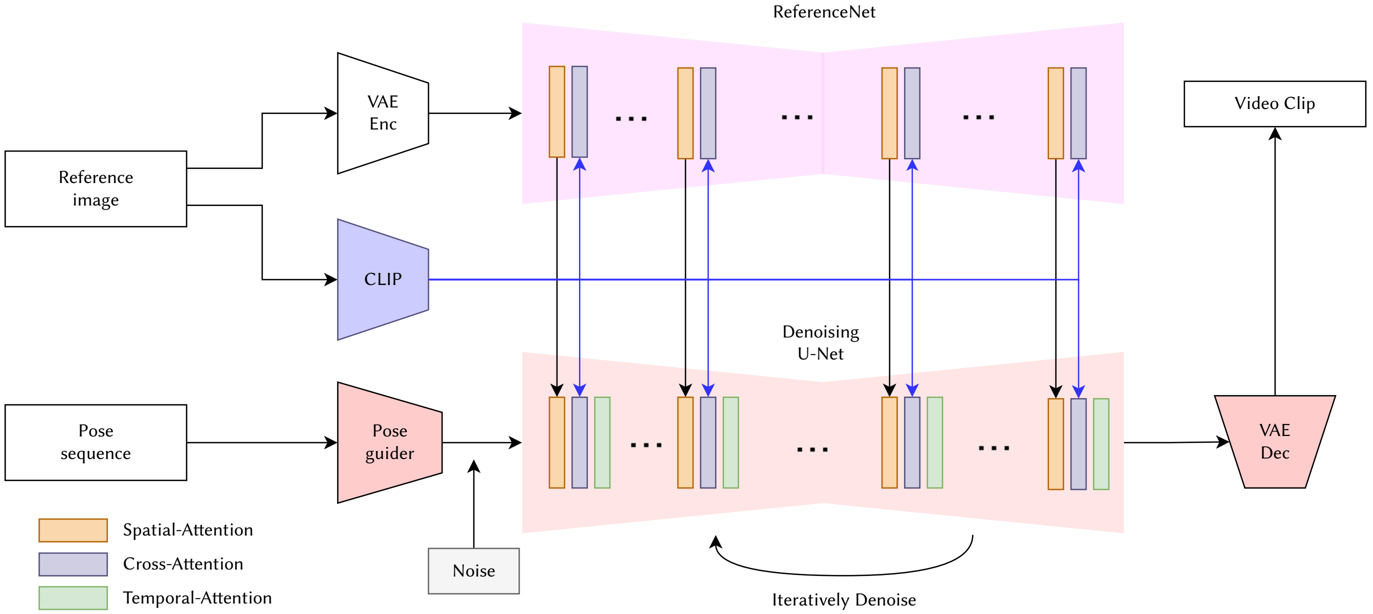


Fig. 8. Animate Anyone [129] architecture.

a novel framework using diffusion models for character animation, see Fig. 8, addressing the challenges of maintaining temporal consistency with detailed character information in image-to-video synthesis. They designed ReferenceNet to merge intricate appearance features from a reference image via spatial attention, and introduced a Pose Guider to ensure controllability and continuity in character movements, along with an effective temporal modeling approach for smooth inter-frame transitions. The method extends Stable Diffusion (SD) by reducing computational complexity through latent space modeling and includes an autoencoder. The network architecture includes ReferenceNet for appearance feature extraction, Pose Guider for motion control, and a temporal layer for continuity of motion. The training strategy consists of two stages: first, training on individual video frames without the temporal layer, and second, introducing and training the temporal layer using a 24-frame video clip. Despite its advancements, the model faces limitations in generating stable hand movements, handling unseen parts during character movement, and operational efficiency due to DDPM. Moreover, Xu *et al.* [130] proposed MagicAnimate, a novel diffusion-based human image animation framework that integrates temporal consistency modeling, precise appearance encoding, and temporal video fusion to synthesize temporally consistent human animation of arbitrary length. They address the challenges of existing methods, which struggle with maintaining temporal consistency and preserving reference identity, by developing a video diffusion model that encodes temporal information with temporal attention blocks and an innovative appearance encoder that retains intricate details of the reference image. MagicAnimate employs a simple video fusion technique to ensure smooth transitions in long animations by averaging overlapping frames. The framework processes animations segment-by-segment to manage memory constraints while leveraging a sliding window method to improve transition smoothness and consistency across segments. This comprehensive approach enables MagicAnimate to produce high-fidelity, temporally consistent animations that faithfully preserve the appearance of the reference image throughout the entire video.

In cases where no motion clue is provided, videos are generated stochastically, constrained solely by the spatial information in the input image. Dorkenwald *et al.* [131] proposed an approach to I2V synthesis by framing it as an invertible domain transfer problem implemented through a Conditional invertible neural network (cINN). To bridge the domain gap between images and videos, they introduced a probabilistic residual representation, ensuring that only complementary information to the initial image is captured. The method allows sampling and synthesizing novel future video progressions from the same start frame. They utilized a separate conditional variational encoder-decoder to compute a compact video representation, facilitating the learning process. Their model captures the interplay between images and videos, explaining video dynamics with a single image and residual information, and supports controlled video synthesis by incorporating additional factors such as motion direction. However, this kind of stochastic video generation can only handle short dynamic patterns in the distribution. Ni *et al.* [132] proposed a method for cI2V generation that synthesizes videos from a single image and a given condition, such as an action label. They introduced Latent flow diffusion models (LFDM), which generate an optical flow sequence in the latent space to warp the initial image, thereby improving the preservation of spatial details and motion continuity. The method involves a two-stage training process: an unsupervised Latent flow auto-encoder (LFAE) to estimate latent optical flow between video frames, and a conditional 3D U-Net-based Diffusion model (DM) to produce temporally-coherent latent flow sequences based on the image and condition. During inference, the image is encoded to a latent map, the condition to an embedding, and the trained DM generates latent flow and occlusion map sequences. During inference, the image is encoded to a latent map, the condition to an embedding, and the trained DM generates latent flow and occlusion map sequences. These sequences warp the latent map to create a new latent map sequence, which is then decoded into video frames. The proposed method, with its decoupled training strategy and efficient operation in a low-dimensional latent flow space, reduces computational cost and complexity while ensuring easy adaptation to new domains.

Wang *et al.* [133] proposed a high-fidelity image-to-video generation method, named DreamVideo, which addresses issues of low fidelity and flickering in existing methods by employing a frame retention branch in a pre-trained video diffusion model. The approach preserves image details by perceiving the reference image through convolution layers and integrating these features with noisy latents. The model incorporates double-condition classifier-free guidance, allowing a single image to generate videos of different actions through varying prompts, enhancing controllable video generation. DreamVideo's architecture includes a primary T2V model and an Image Retention block that infuses image control signals into the U-Net structure. During inference, the model combines text and image inputs to generate contextually consistent videos using CLIP text embeddings and a U-Net-based generative process. Additionally, the Two-Stage Inference method extends video length and creates varied content by using the final frame of one video as the initial frame for the next, showcasing the model's strong image retention and video generation capabilities. Zhang *et al.* [134] proposed I2VGen-XL, a method utilizing two stages of cascaded diffusion models to achieve high semantic consistency and spatiotemporal continuity in video synthesis. The approach addresses challenges in semantic accuracy, clarity, and continuity by decoupling semantic and qualitative factors, using static images as guidance. The base stage ensures semantic coherence and preserves content at low resolution with two hierarchical encoders—a fixed CLIP encoder for high-level semantics and a learnable content encoder for low-level details. The refinement stage enhances video resolution and refines details using a brief text input and a separate video diffusion model. Training involves initializing the base model with pre-trained SD2.1 parameters and moderated updates, while the refinement model undergoes high-resolution training and fine-tuning on high-quality videos. Inference employs a noising-denoising process and DDIM/DPM-solver++ to generate high-resolution videos from low-resolution outputs.

To create more controllable videos, various motion cues like predefined directions and action labels are used. Blattmann *et al.* [135] proposed an approach for generating videos from static images by learning natural object dynamics through local pixel manipulations. Their generative model learns from videos of moving objects without needing explicit information about physical manipulations and infers object dynamics in response to user interactions, understanding the relationships between different object parts. The goal is to predict object deformation over time from a static image and a local pixel shift, using two encoding functions: an object encoder for the current object state and an interaction encoder for the pixel shift. They utilize a hierarchical recurrent model to understand complex object dynamics, predicting a sequence of object states in response to the pixel shift. Object dynamics are modeled using a flexible prediction function based on Recurrent Neural Networks (RNN), with higher-order dynamics captured by introducing a hierarchy of RNN predictors operating on different spatial scales. The decoder generates individual image frames from the predicted object states using a hierarchical image-to-sequence UNet structure. Instead of ground-truth interactions, dense optical flow displacement maps are used to simulate training pokes, minimizing the perceptual distance between predicted and actual video frames. Training involves pretraining the encoders and decoder to reconstruct image frames, then refining the model to predict object states and synthesize video sequences. Their interactive I2V synthesis model allows users to specify the desired motion through the manual poking of a pixel.

In addition, Menapace *et al.* [136] proposed a novel framework for the Playable video generation (PVG) task, which generates videos from the first frame and a sequence of discrete actions. While the PVG task reduces user burden by not requiring detailed motion information, it struggles with generating videos involving complex motions. An unsupervised learning approach is adopted that allows users to control video generation by selecting discrete actions at each time step, similar to video games. The framework, named Clustering for Action Decomposition and DiscoverY (CADDY), learns semantically consistent actions and generates realistic videos based on user input using a self-supervised encoder-decoder architecture driven by a reconstruction loss on the generated video. CADDY discovers distinct actions via clustering during the generation process, employing an encoder-decoder with a discrete bottleneck layer to capture frame transitions without needing action label supervision or a predefined number of actions. The action network estimates action label posterior distributions by decomposing actions into discrete labels and continuous components, ensuring meaningful action labels by preventing direct encoding of environment changes in the variability embeddings.

Within the generation of dynamic videos from static images presents a trend very similar to the previous section, Text-to-Video Synthesis, where we can see how attention mechanisms, autoencoders and diffusion models stand out. As we can see, GANs are not as frequent as in synthetic image generation. This approach to video generation can raise more ethical concerns than the previous one, as it can use images of real people and generate videos that can potentially harm them; whereas in the previous section, it involves content generated completely from scratch.

3. Video-to-Video Synthesis

Video-to-video (V2V) synthesis is an advanced field focused on generating realistic video sequences by transforming or translating visual information from one video domain to another. The main goal is to create high-quality, temporally consistent videos that adhere to specific input conditions, such as text, pose, style, or semantic maps. Recent advancements in this area have introduced several techniques to enhance the quality, efficiency, and consistency of video synthesis, thus pushing the boundaries of what is possible in video generation. Wang *et al.* [137] proposed a three-stage framework for human pose transfer in videos, focusing on transferring dance poses from a source person in one video to a target person in another. The process begins with the extraction of frames and pose masks from both source and target videos. Subsequently, a model synthesizes frames of the target person in the desired dance pose, followed by a refinement phase to enhance the quality of these frames. The model comprises several key components, including pose extraction and normalization, a GAN-based synthesis using Cross-domain correspondence network (CoCosNet), and a coarse-to-fine strategy with two GANs for detailed face reconstruction and smooth frame sequences. Their approach involves visualizing keypoints to create pose skeleton labels, adjusting for differences in body proportions, learning the translation from pose domain to image domain, and matching features for coherent synthesis. Although their method outperforms existing approaches, it still encounters challenges with large pose variations and domain generalization, which suggests potential areas for future improvement.

Furthermore, Zhuo *et al.* [138] introduced Fast-Vid2Vid, a spatial-temporal compression framework designed to reduce computational costs and accelerate inference in Video-to-Video synthesis (Vid2Vid). While traditional Vid2Vid generates photorealistic videos from semantic maps, it suffers from high computational costs due to the network architecture and sequential data streams. Zhuo *et al.* addressed this by introducing Motion-aware inference (MAI) to compress the input data stream without altering network parameters and developing Spatial-temporal knowledge distillation (STKD) to transfer knowledge from a high-resolution teacher model to a low-resolution student model. Their approach incorporates Spatial knowledge distillation (Spatial KD) for generating high-resolution frames from low-resolution inputs and Temporal knowledge distillation (Temporal KD) to maintain temporal coherence in sparse video sequences. Additionally, they utilize a part-time student generator for sparse frame synthesis and a fast motion compensation method for interpolating intermediate frames, thereby reducing computational load while maintaining visual quality. Further advancing the field, Yang *et al.* [139] introduced a zero-shot text-guided video-to-video translation framework that adapts image models for video applications. This framework is composed of key frame translation and full video translation. Key frames are generated using an adapted diffusion model with hierarchical cross-frame constraints to ensure coherence in shapes, textures, and colors. These frames are then propagated to the rest of the video using temporal-aware patch matching and frame blending, achieving both global style and local texture temporal consistency without requiring re-training or optimization. A key innovation of this approach is the use of optical flow for dense cross-frame constraints, ensuring consistency across different stages of diffusion sampling. However, the method's reliance on accurate optical flow can lead to artifacts if the flow is incorrect, and significant appearance changes may disrupt temporal consistency, limiting the ability to create unseen content without user intervention.

Following the trend of previous researchers but focusing on zero-shot techniques, Wang *et al.* [140] presented vid2vid-zero, a zero-shot video editing method that leverages pre-trained image diffusion models without requiring video-specific training. Their method introduces a null-text inversion module for text-to-video alignment, a cross-frame modeling module for temporal consistency, and a spatial regularization module to preserve the fidelity of the original video. Vid2vid-zero addresses the issue of flickering in frame-wise image editing by ensuring temporal consistency through a Spatial-temporal attention (ST-Attn) mechanism, which balances bi-directional temporal information and spatial alignment using pre-trained diffusion models. While effective in video editing tasks, the method's reliance on pre-trained image models limits its capacity to edit actions in videos due to the absence of temporal and motion priors. Expanding on the idea of zero-shot video editing, Qi *et al.* [141] proposed FateZero, a zero-shot text-based editing method for real-world videos that does not require per-prompt training or user-specific masks. To achieve consistent video editing, FateZero utilizes techniques based on pre-trained models, capturing intermediate attention maps during DDIM inversion to retain structural and motion information and fusing these maps during editing. A blending mask, derived from cross-attention features, minimizes semantic leakage, while the reformed self-attention mechanism in the denoising UNet enhances frame consistency. Despite its impressive performance, FateZero faces challenges in generating entirely new motions or significantly altering shapes.

Other authors have opted for the use of diffusion models, due to their performance in similar tasks. Molad *et al.* [142] proposed Dreamix, a text-driven video editing method that uses a text-conditioned video diffusion model (VDM). Dreamix preserves the original video's fidelity by initializing with a degraded version of the input video and then fine-tuning the model. This mixed fine-tuning technique enhances motion editability by incorporating individual frames with masked temporal attention. Dreamix achieves text-guided video editing by inverting corruptions, downsampling the input video, corrupting it with noise, and then upscaling it using cascaded diffusion models aligned with the text prompt. This approach effectively preserves low-resolution details while synthesizing high-resolution outputs. Focusing on motion guidance, Hu *et al.* [143] introduced VideoControlNet, a motion-guided video-to-video translation framework using a diffusion model with ControlNet. Inspired by video codecs, VideoControlNet leverages motion information to maintain content consistency and prevent redundant regeneration. The first frame (I-frame) is generated using the diffusion model with ControlNet, mirroring the structure of the input frame. Key frames (P-frames) are then generated using the motion-guided P-frame generation (MgPG) module, which employs motion information for consistency and inpaints occluded areas using the diffusion model. The remaining frames (B-frames) are efficiently interpolated using the motion-guided B-frame interpolation (MgBI) module. This framework produces high-quality, consistent videos by utilizing advanced inpainting methods alongside motion information.

Adding to the discussion of temporal consistency, Liang *et al.* [144] introduced FlowVid, a V2V synthesis framework that ensures temporal consistency across frames by leveraging spatial conditions and temporal optical flow clues from the source video. Unlike previous methods, FlowVid uses optical flow as a supplementary reference to handle imperfections in flow estimation. The model warps optical flow from the first frame and uses it in a diffusion model, enabling the propagation of edits made to the first frame throughout subsequent frames. FlowVid extends the U-Net architecture to include a temporal dimension and is trained using joint spatial-temporal conditions, such as depth maps and flow-warped videos, to maintain frame consistency. During generation, the model edits the first frame with prevalent Image-to-image (I2I) models and propagates these edits using a trained model, incorporating global color calibration and self-attention feature integration to preserve structure and motion, thus achieving effective video synthesis with high temporal consistency. In a similar pursuit of enhancing temporal coherence, Wu *et al.* [145] proposed Fairy, a minimalist yet robust adaptation of image-editing diffusion models for video editing. Fairy improves temporal consistency and synthesis fidelity through anchor-based cross-frame attention, which propagates diffusion features across frames. To handle affine transformations, Fairy employs a unique data augmentation strategy, enhancing the model's equivariance and consistency. The anchor-based model samples K anchor frames to extract and propagate diffusion features, ensuring consistency by aligning similar semantic regions across frames. While Fairy excels in maintaining temporal consistency, its strong focus on this aspect reduces its accuracy in rendering dynamic visual effects, such as lightning or flames.

Lastly, several other methods offer significant contributions to the video-to-video synthesis domain. Ku *et al.* [146] proposed AnyV2V, see Fig. 9, a training-free video editing framework that simplifies video editing into two steps: editing the first frame with any image editing model and using an image-to-video generation model to create the edited video through temporal feature injection. AnyV2V is compatible with various image editing tools, allowing for diverse edits such as style transfer, subject-driven editing, and identity manipulation, without the need for fine-tuning. The framework uses DDIM inversion for structural guidance and feature injection to maintain consistency in appearance and motion, enabling accurate and flexible video editing. Additionally, it supports long video editing by handling videos beyond the training frame lengths of current I2V models, outperforming existing methods in user evaluations and standard metrics. Ouyang *et al.* [147] introduced I2VEdit, a video editing solution designed to extend the capabilities of image editing tools to videos. This approach achieves this by propagating single-frame edits throughout an entire video using a pre-trained Image-to-video model. Notably, I2VEdit adapts to the extent of edits, preserving visual and motion integrity while handling various types of edits, including global, local, and moderate shape changes. The method's core processes, coarse motion extraction and appearance refinement, play crucial roles in ensuring consistency. Coarse motion extraction captures basic motion patterns through a motion LoRA and employs skip-interval cross-attention to mitigate quality degradation in long videos.

Meanwhile, appearance refinement uses fine-grained attention matching for precise adjustments and incorporates Smooth area random perturbation (SARP) to enhance inversion sampling. To achieve its results, I2VEdit segments the source video into clips, processes each clip for motion and appearance consistency, and refines appearances using EDM [99] inversion and attention matching. Building on this, Ouyang *et al.* [148] further proposed Content deformation field (CoDeF), a novel video representation, emphasizing its application in Video-to-video translation. CoDeF introduces a canonical content field for static content aggregation and a temporal deformation field for recording

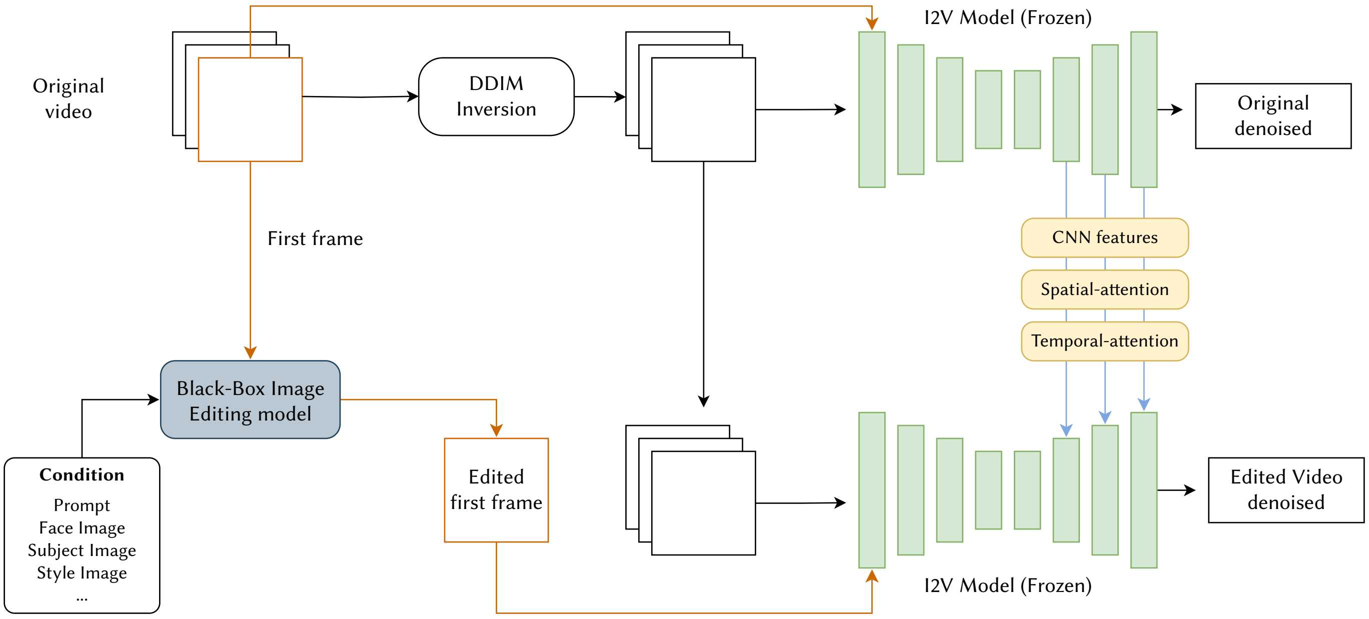


Fig. 9. AnyV2V [146] framework.

frame transformations. This approach optimizes the reconstruction of videos while preserving essential semantic details, such as object shapes. In the context of Video-to-video translation, CoDeF employs ControlNet on the canonical image, which significantly enhances temporal consistency and texture quality compared to state-of-the-art zero-shot video translations using generative models. By avoiding the need for time-intensive inference models, this process becomes more efficient. The canonical image, optimized through CoDeF, serves as a basis for applying image algorithms, ensuring consistent effect propagation across the entire video via the temporal deformation field.

A different approach to video editing with VideoSwap was presented by Gu *et al.* [149], focusing on customized video subject swapping. Unlike methods relying on dense correspondences, VideoSwap utilizes semantic point correspondences, allowing the replacement of the main subject in a video with a target subject of a different shape and identity, all while preserving the original background. The approach includes encoding the source video, applying DDIM inversion, and using semantic points to guide the subject's motion trajectory. The process also involves extracting and embedding semantic points, registering these points for motion guidance, and enabling user interactions to refine motion and shape alignment. Recently, Bai *et al.* [150] proposed UniEdit, a tuning-free framework for video motion and appearance editing. This framework leverages a pre-trained Text-to-video generator in an inversion-then-generation pipeline. UniEdit addresses content preservation by using temporal and spatial self-attention layers to encode inter-frame and intra-frame dependencies. Additionally, it introduces auxiliary reconstruction and motion-reference branches to inject the desired source and motion features into the main editing path. For content preservation, the auxiliary reconstruction branch injects attention features into the spatial self-attention layers. Motion injection, on the other hand, is achieved by guiding the main path with a motion-reference branch during denoising, utilizing temporal attention maps for alignment with the target prompt. In appearance editing, UniEdit maintains structural consistency by implementing spatial structure control while omitting the motion-reference branch. Despite its robust capabilities, UniEdit faces challenges, particularly when addressing motion and appearance editing simultaneously.

In this section we have analyzed the latest research related to video-to-video synthesis. Within this field we have seen how the most used technique is the diffusion model, as we have seen in other sections of this survey. This is in line with expectations due to all the attention they are receiving in recent years. However, we can also see that other methodologies such as GANs or attention mechanisms are also used. We have also noted that several papers use a zero-shot approach to address the problem.

4. Text-Image-to-Video Synthesis

Text-image-video synthesis (TI2V) is a growing field of research focused on generating dynamic video content from static images and text descriptions. Given a single image I and text prompt T, text-image-to-video generation aims to synthesize I new frames to yield a realistic video, $I = \langle I^0, I^1, ..., I^M \rangle$ y starting from the given frame I^0 and satisfying the text description T . This field aims to bridge the gap between different modalities to create coherent and contextually accurate videos. Several approaches have been developed to address the challenges in this domain, ranging from aligning visual and textual information to ensuring temporal consistency and control over generated content. Hu *et al.* [151] proposed a novel video generation task called Text-Image-to-Video (TI2V) generation, which creates videos from a static image and a text description, focusing on controllable appearance and motion. They introduced the Motion Anchor-based video GEnerator (MAGE) to address key challenges such as aligning appearance and motion from different modalities and handling text description uncertainties. MAGE uses a Motion anchor (MA) structure to store aligned appearance-motion representations and incorporates explicit conditions and implicit randomness to enhance diversity and control. The framework employs a VQ-VAE encoder-decoder architecture for visual token representation and uses three-dimensional axial transformers to recursively generate frames. Training involves a supervised learning approach to approximate the conditional distribution of video frames based on the initial image and text. The motion anchor aligns text-described motion with visual features, ensuring consistent and diverse video output through auto-regressive frame generation.

Complementing this, Guo *et al.* [152] proposed AnimateDiff, a practical framework for animating personalized T2I models without requiring model-specific tuning. The core of the framework is a plug-and-play motion module, trained to learn transferable motion priors from real-world videos, which can be integrated into any personalized T2I model. The training process involves three stages: fine-tuning a domain adapter to align with the target video dataset, introducing and optimizing a motion module for motion modeling, and using MotionLoRA, a lightweight fine-tuning technique, to adapt the pre-trained motion module to new motion patterns with minimal data and training cost. AnimateDiff effectively addresses the problem of animating personalized T2Is while preserving their visual quality and domain knowledge, demonstrating the adequacy of Transformer architecture for modeling motion priors and offering an efficient solution for users who desire specific motion effects without bearing the high costs of pre-training. In contrast, Yin *et al.* [153] proposed NUWA-XL, a novel "Diffusion over Diffusion" architecture for generating extremely long videos. Unlike traditional methods that generate videos sequentially, leading to inefficiencies and a training-inference gap, NUWA-XL uses a "coarse-to-fine" process where a global diffusion model generates keyframes and local models fill in between, allowing parallel generation. The architecture incorporates Temporal KLVAE to compress videos into low-dimensional latent representations and Mask temporal diffusion (MTD) to handle both global and local diffusion processes using masked frames. Although NUWA-XL is currently validated on cartoon data due to the lack of open-domain long video datasets, it shows promise in overcoming data challenges and improving efficiency, albeit requiring substantial GPU resources for parallel inference.

Esser *et al.* [154] proposed a structure and content-guided video diffusion model that edits videos based on user descriptions. They resolved conflicts between content and structure by training on monocular depth estimates with varying detail levels and introduced a novel guidance method for temporal consistency through joint video and image training. The approach extends latent diffusion models to video by incorporating temporal layers into a pre-trained image model, adding 1D convolutions and self-attentions to residual and transformer blocks. The encoder downsamples images to a latent code, improving efficiency, while depth maps and CLIP embeddings are used for structure and content conditioning, respectively. This approach allows full control over temporal, content, and structure consistency without requiring per-video training or pre-processing, showing improved temporal stability and user preference over related methods. Expanding on the concept of control, Yin *et al.* [155] proposed DragNUWA, an open-domain diffusion-based video generation model that integrates text, image, and trajectory inputs to provide fine-grained control over video content from semantic, spatial, and temporal perspectives. They address the limitations of current methods, which focus on only one type of control and struggle with complex trajectory handling, by introducing advanced trajectory modeling techniques: a Trajectory sampler (TS) for arbitrary trajectories, Multiscale fusion (MF) for controlling trajectories at different granularities, and an Adaptive training (AT) strategy for generating consistent videos. DragNUWA can generate realistic and contextually consistent videos by leveraging the combined inputs of text, images, and trajectories during both training and inference.

Further enhancing controllability, Wang *et al.* [156] proposed VideoComposer, a system for enhancing controllability in video synthesis through the use of temporal conditions like motion vectors. They introduced a Spatio-temporal condition encoder (STC-encoder) to integrate spatial and temporal dependencies, ensuring inter-frame consistency. The system decomposes videos into textual, spatial, and temporal conditions, and uses a latent diffusion model to recompose videos based on these inputs. Textual conditions provide coarse-grained visual content, while spatial conditions offer structural and stylistic guidance. Temporal conditions, including motion vectors and depth sequences, allow detailed control of temporal dynamics. Recently, Ni *et al.* [157] proposed TI2V-Zero, a zero-shot, tuning-free method for text-conditioned Image-to-video (TI2V) generation that leverages a pretrained T2V diffusion model. This approach avoids costly training, fine-tuning, or additional modules by using a "repeat-and-slide" strategy to condition video generation on a provided image, ensuring temporal continuity through a DDPM inversion strategy and resampling techniques. The method uses a 3D-UNet-based denoising network and modulates the reverse denoising process to generate videos frame-by-frame, preserving visual coherence and consistency, thus enabling the synthesis of long videos while maintaining high visual quality.

In this section where we have analyzed the techniques to generate videos from static images and textual descriptions, we have seen again a main focus, which are the diffusion models, i.e. a trend is observed, which seems to show that it will be the most used technique in the coming years. In addition, we also continue to observe other approaches such as attention mechanisms or autoencoders. The greatest danger of this set of techniques, like the previous one, is that they can use images of people to create complete videos, which can cause serious damage. However, not all applications of these techniques are negative.

5. Multi-Modal Video Generation

Multi-Modal Video Generation (MMVG) refers to a versatile field in which video content is synthesized based on different forms of input, such as text, images, or existing videos. Although models like Sora and Genie can accept various types of input, they typically process one modality at a time—either generating videos from text descriptions, animating static images, or transforming existing video footage. These approaches leverages the strengths of different data modalities to produce highly realistic and contextually coherent videos. The core objective of MMVG is to create coherent, high-fidelity, temporal consistent videos by leveraging the strengths of each input type. Recent advancements in this field have led to the development of sophisticated models capable of interpreting and synthesizing complex scenes by concurrently analyzing textual descriptions, visual cues, and pre-existing video footage. These models push the boundaries of video generation, offering versatile applications in content creation, entertainment, and beyond.

More recently, *OpenAI* [6] introduced Sora, a diffusion model that represents a significant advancement in T2V generation by training a model from scratch rather than fine-tuning pre-trained models. Drawing from transformer architecture scalability, Sora replaces the conventional U-Net with a transformer-based structure, effectively managing large-scale video data for complex generative tasks. Sora can generate high-fidelity videos up to a minute long, maintaining visual quality and narrative consistency across multiple shots. It leverages a patch-based approach, turning visual data into spacetime patches, which enhances its ability to handle videos and images of varying durations, resolutions, and aspect ratios. Sora excels in linguistic comprehension, accurately following detailed prompts to generate coherent video content. However, it faces challenges in rendering realistic interactions and comprehending complex scenes with multiple active elements. Despite these limitations, Sora's capabilities in video-to-video editing, image animation, and extending generated videos mark a significant step toward building general-purpose simulators of the physical world. Bruce *et al.* [7] introduced Genie, a generative interactive environment model trained unsupervised from unlabelled Internet videos. Genie uses spatiotemporal transformers, a novel video tokenizer, and a causal action model to create diverse, action-controllable virtual worlds from various inputs such as text, images, and sketches. It generates video frames autoregressively, enabling interaction on a frame-by-frame basis without ground-truth action labels.

TABLE IV. Comprehensive Overview of a Few Synthetic Video Generation Techniques

Models	Year	Technique	Target Outcome	Data Used	Open Source
Make-A-Video [91]	2023	Transformer-based	Text-to-video synthesis	Various	No
Video Diffusion [89]	2023	Diffusion-based	High-quality video synthesis	Video datasets	No
VideoPoet [5]	2023	Transformer-based	Generate poetic video narratives	Web-collected dataset	No
Godiva [104]	2023	GAN-based	Generate dynamic video content	High-resolution video datasets	No
CogVideo [106]	2023	Transformer-based	Extend CogView into video	Diverse text and video datasets	Yes
NUWA [107]	2023	Transformer-based	Synthesize coherent video clips	Diverse content from web datasets	No
NUWA-Infinity [108]	2023	Transformer-based	Generate endless video streams	Extended NUWA dataset	No
VideoGPT [109]	2023	GPT-based	Utilize GPT architecture	Various video datasets	Yes
Video LDMs [110]	2024	Latent Diffusion Models	Implement latent space techniques	Various	No
Text-to-Video (T2V) [158]	2023	Transformer-based	Synthesize video from static images	Diverse image and video datasets	No
ModelScope Text-to-Video [113]	2024	Transformer-based	Scalable text-to-video model	Large-scale web-collected video datasets	Yes
W.A.L.T [114]	2023	Diffusion Models	Enhance video synthesis	Various	No
C-ViViT [115]	2023	VAE-based	Create detailed videos from categories	Category-labeled video datasets	No
Text2Video-Zero [118]	2023	Zero-Shot Learning	Generate videos without explicit training	General video datasets	Yes
ART•V [119]	2024	AI Rendered Textures	Artistic video creation	Artistic style datasets	No
BIVDiff [120]	2023	Bi-directional Diffusion	Bidirectional control over video generation	Various	Yes
Simple Diffusion Adapter [116]	2024	Diffusion Models	Simplify diffusion processes	Various	Yes
HiGen [121]	2024	Hierarchical Generation	Layered approach to video scenes	Multi-layer video datasets	Yes

TABLE V. Overview of Techniques for Detecting AI-Generated Videos

Authors	Year	Technique	Target Outcome	Data Used	Open Source
Vahdati *et al.* [159]	2024	Synthetic video detection by forensic trace analysis	Detect AI-generated synthetic videos	Synth-vid-detect	No
He *et al.* [160]	2024	Temporal defects analysis	Identify temporal defects in AI-generated videos	ExposingAI-Video	No
Chen *et al.* [162]	2024	Detail Mamba for spatial-temporal artifacts detection	Enhance detection of AI-generated videos	GenVideo	Yes
Bai *et al.* [163]	2024	Spatio-temporal CNN analysis	Detect AI-generated videos using motion discrepancies	GVD	Yes
Ma *et al.* [164]	2024	Temporal artifact focus	Focus on temporal artifacts in video detection	GVF	Yes
Ji *et al.* [165]	2024	Dual-Branch 3D Transformer	Integrate motion and visual appearance for fake video detection	GenVidDet	No
Liu *et al.* [167]	2024	Diffusion-generated video detection	Capture spatial and temporal features in RGB frames and DIRE values	TOINR	No

As we can see, this section, multimodal video generation, is the least explored of all the approaches analyzed, see Table IV, and possibly the most complex, since we not only have to generate the visual part of the videos, but also the audio. In addition, we must ensure that both are matched and do not generate easily detectable artifacts. The techniques analyzed in this field are diffusion models and transformers. Possibly this area will be explored in more detail in the coming years.

B. Detection of AI-Generated Videos

In the rapidly evolving landscape of Generative AI (Gen AI), significant progress has been made in developing techniques to detect AI-generated synthetic images. Given that a video can be viewed as a sequence of images, one might reasonably expect that synthetic image detectors would also be effective at identifying AI-generated synthetic videos. Surprisingly, Vahdati *et al.* [159] reveal that current synthetic image detectors fail to reliably detect synthetic videos. Their study demonstrates that the forensic traces left by synthetic video generators are markedly different from those produced by image generators. This issue is not due to the degradation effects of H.264 compression but rather to the distinct characteristics of video generation. Therefore, their findings underscore the urgent need for detection methods tailored specifically to synthetic video content. Table V provides an overview of the techniques used for detecting AI-generated videos, highlighting key approaches and their application to various datasets. Despite the growing concerns, research into detecting synthetic videos has been relatively limited. Video generation technology is still in its early stages compared to image generation, and as a result, fewer detection methods are available. However, recent efforts have started to address this gap (see Fig. 10).

One early approach comes from, He *et al.* [160] who proposed a novel detection method for identifying AI-generated videos by analyzing temporal defects at both local and global levels. The method is based on the assumption that AI-generated videos exhibit different temporal dependencies compared to real videos due to their distinct capturing

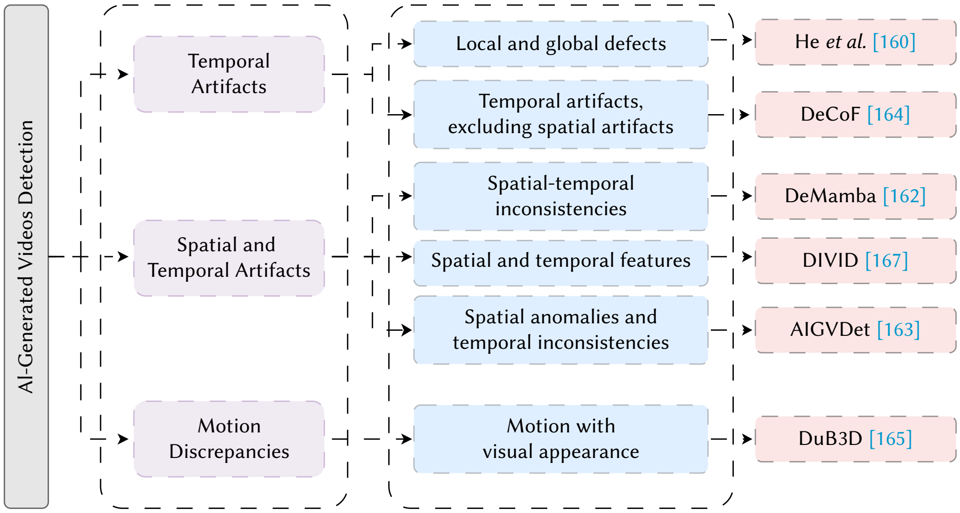

AI-Generated Videos Detection
Temporal Artifacts
Spatial and Temporal Artifacts
Motion Discrepancies
Local and global defects
Temporal artifacts, excluding spatial artifacts
Spatial-temporal inconsistencies
Spatial and temporal features
Spatial anomalies and temporal inconsistencies
Motion with visual appearance
He et al. [160]
DeCoF [164]
DeMamba [162]
DIVID [167]
AIGVDet [163]
DuB3D [165]

Fig. 10. AI-Generated videos detection methods overview.

and generation processes. Real videos, which are captured by cameras, have high temporal redundancy, whereas AI-generated videos control frame continuity in the latent space, leading to defects at different spatio-temporal scales. To address local motion information, the method uses a frame predictor trained on real videos to measure inter-frame motion predictability. Fake videos show larger prediction errors because they have less temporal redundancy. Temporal aggregation is employed to maintain long-range information and reduce the impact of diverse spatio-temporal details. The aggregated error map is then processed by a 2D encoder to obtain local motion features. For global appearance variation, the method extracts visual features using a pre-trained BEiT v2 [161] image encoder. These features are fed into a transformer to model temporal variations, identifying abnormal appearance changes across frames. Finally, a channel attention-based fusion module combines the local motion and global appearance features to enhance detection reliability. This module adjusts channel significance to extract more generalized forensic clues.

Furthermore, Chen *et al.* [162] proposed a plug-and-play module named Detail Mamba (DeMamba), designed to enhance the detection of AI-generated videos by identifying spatial and temporal artifacts. DeMamba builds upon the Mamba framework to explore both local and global spatial-temporal inconsistencies, addressing the limitation of models that consider only one aspect, either spatial or temporal. Using vision encoders like CLIP and XCLIP, it encodes video frames into a sequence of features, groups them spatially, and applies the DeMamba module to model intra-group consistency. Aggregated features from different groups help determine video authenticity. The DeMamba module introduces a novel approach to spatial consolidation by splitting features into zones along height and width, performing a 3D scan for spatial-temporal input. Unlike previous mechanisms, DeMamba's continuous scan aligns spatial tokens sequentially, enhancing the model's ability to capture complex relationships. For classification, DeMamba averages input features to obtain global features and pools processed features into local features, concatenating them with the global ones for classification via a simple MLP, ensuring robust video authenticity detection.

Based on the assumption that low-quality videos show abnormal textures and physical rule violations, while high-quality videos, indistinguishable to the naked eye, often manifest temporal discontinuities in optical flow maps, Bai *et al.* [163] proposed an effective AI-generated video detection (AIGVDet) scheme by capturing forensic traces with a two-branch spatio-temporal Convolutional Neural Network (CNN). This scheme employs two ResNet sub-detectors to identify anomalies in the spatial and optical flow domains. The spatial detector examines the abnormality of spatial pixel distributions within single RGB frames, while the optical flow detector captures temporal inconsistencies via optical flow. The model uses RGB frames and optical flow maps as inputs, with the two-branch ResNet50 encoder detecting abnormalities and a decision-level fusion binary classifier combining this information for the final prediction. AIGVDet effectively leverages motion discrepancies for comprehensive spatio-temporal analysis to detect AI-generated videos. Ma *et al.* [164] found that detectors based on spatial artifacts lack generalizability. Hence, they proposed DeCoF, a detection model that focuses on temporal artifacts and eliminates the impact of spatial artifacts during feature learning. DeCoF is the first method to use temporal artifacts by decoupling them from spatial artifacts, mapping video frames to a feature space where inter-feature distance is inversely correlated with image similarity, and detecting anomalies from inter-frame inconsistency. The method reduces computational complexity and memory requirements, needing only to learn anomalies between features. However, DeCoF may experience significant performance degradation or be inapplicable in the face of tampered video, such as Deepfake and malicious editing.

Traditional video detection models often overlook specific characteristics of downstream tasks, particularly in fake video detection where motion discrepancies between real and generated videos are significant, as generators tend to excel in appearance modeling but struggle with accurate motion representation. Ji *et al.* [165] proposed the Dual-Branch 3D Transformer (DuB3D) to address this issue by integrating motion information with visual appearance using a dual-branch architecture that fuses raw spatio-temporal data and optical flow. The spatial-temporal branch processes original frames to capture spatial-temporal information and identify anomalies, while the optical flow branch uses GMFlow [166] to estimate and capture motion information, and these features are combined using a Multi-layer perceptron (MLP) for classification. Built on the Video Swin Transformer backbone, DuB3D effectively enhances fake video detection by emphasizing motion modeling and demonstrating strong generalization across various video types. More recently, Liu *et al.* [167] proposed a novel approach for DIffusion-generated VIdeo Detection (DIVID). DIVID uses CNN+LSTM architectures to capture both spatial and temporal features in RGB frames and DIRE values. Initially, the CNN is fine-tuned on original RGB frames and DIRE values, followed by training the LSTM network based on the CNN's feature extraction. This two-phase training enhances detection accuracy for both in-

TABLE VI. AI-Generated Image Detection Datasets

Dataset	Year	Content	Real Source	Generator	#Real	#Generated	Available
LSUN Bed [168]	2022	Bedroom	LSUN	GAN/DM	420,000	510,000	✓
DFF [169]	2023	Face	IMDB-WIKI	DM	30,000	90,000	✓
RealFaces [170]	2023	Face	-	DM	-	25,800	✓
DiffusionForensics [79]	2023	General	LSUN ImageNet	DM	134,000	481,200	✓
Synthbuster [76]	2023	General	Raise-1k	DM	-	9,000	✓
DDDB [171]	2023	Art	LAION-5B	DM	64,479	73,411	✓
DE-FAKE [172]	2023	General	MSCOCO Flickr30k	DM	-	191 946	✗
AI-Gen [173]	2023	General	ALASKA	DM	20,000	40,000	✗
ArtiFact [174]	2023	General	Various sources including AFHQ, CelebAHQ, COCO, etc.	GAN/DM	964,989	1,531,749	✓
AutoSplice [175]	2023	General	Visual News	DM	2,273	3,621	✓
HiFi-IFDL [176]	2023	General	Various sources including AFHQ, CelebAHQ, LSUN, Youtube face etc.	GAN/DM	~ 600,000	1,300,000	✓
M3DSYNTH [177]	2023	CT	LIDC-IDRI	GAN/DM	1,018	8,577	✓
DIF [74]	2023	General	Laion-5B	GAN/DM	168,600	168,600	✓
DGM^4 [178]	2023	General	News: The Guardian, BBC, USA TODAY, Washington Post	GAN/DM	77,426	152,574	✓
COCOFake [179]	2023	General	COCO	DM	~ 1,200,000	~ 1,200,000	✓
DiFF [180]	2024	Face	VoxCeleb2 CelebA	DM	23,661	537,466	✓
CIFAKE [181]	2024	General	CIFAR-10	DM	60,000	60,000	✓
GenImage [182]	2024	General	ImageNet	GAN/DM	1,331,167	1,350,000	✓
Fake2M [183]	2024	General	CC3M	GAN/DM	-	2,300,000	✓
WildFake [184]	2024	General	Various sources including COCO, FFHQ Laion-5B, etc.	GAN/DM	1,013,446	2,680,867	✗

domain and out-domain videos. Diffusion Reconstruction Error (DIRE) is calculated as the absolute difference between an original image and its reconstructed version from a pre-trained diffusion model, capturing signals of diffusion-generated images. By training the CNN+LSTM with DIRE and RGB frame features, DIVID improves detection accuracy for AI-generated videos.

Detecting AI-generated videos is an emerging challenge, distinct from synthetic image detection due to unique forensic traces in video content. While promising methods have begun to address this gap, leveraging spatio-temporal analysis and novel fusion techniques, the field is still evolving, see Table V. Continued innovation is essential to stay ahead of rapidly advancing video generation technologies.

V. Datasets

One of the most important aspects of DL model development is the availability of quality datasets. These datasets have to have some fundamental properties to be able to create robust models: to be representative, intra-class variability, balance between classes and a minimum quality. This will allow us to create suitable new generative and detection models. In this section we will focus on image and video datasets generated with AI.

The development of AI-generated images relies heavily on the availability of diverse and comprehensive datasets. These datasets provide the essential training material for models to learn from, enabling them to generate realistic and varied images. Ranging from large-scale collections of image-text pairs to datasets specifically designed for detecting synthetic content, these resources play a pivotal role in advancing the field. Regarding detection, we need representative and varied datasets that include different generation techniques and models. This will allow the development of robust models capable of being applied in real situations.

A. Image Datasets

In this section, we highlight some of the key image datasets that have significantly contributed to state-of-the-art AI-generated imagery. These datasets not only differ in size and content but also cater to various research needs, from general-purpose image generation to specialized tasks like AI-generated images detection and multimodal learning. For a detailed comparison, refer to Table VI, which summarizes the features and scope of these datasets.

Conceptual Captions 12M (CC12M) [185] is a large-scale dataset of 12.4 million image-text pairs derived from the Conceptual Captions 3M (CC3M) dataset [186]. CC12M was created by relaxing some of the filters used in CC3M to increase the recall of potentially useful image-alt-text pairs. The relaxed filters allow for more diverse and extensive data, though this results in a slight drop in precision. Unlike CC3M, CC12M does not perform hypernymization or digit substitution, except for substituting person names to protect privacy. This dataset's larger scale and diversity make it well-suited for vision-and-language pre-training tasks.

WIT [187] introduced to facilitate multimodal, multilingual learning, contains 37.5 million entity-rich image-text examples and 11.5 million unique images across 108 Wikipedia languages. It serves as a pre-training dataset for multimodal models, particularly useful for tasks like image-text retrieval. WIT stands out due to its large size, multilingual nature with over 100 languages, diverse concepts, and a challenging real-world test set. It combines high-quality image-text pairs from curated datasets like Flickr30K and MS-COCO with the scalability of extractive datasets. WIT's creation involved filtering low-information associations and ensuring image quality. The dataset provides multiple text types per image (reference, attribution, and alt-text), offers extensive cross-lingual text pairs, and supports contextual understanding with 120 million contextual texts.

RedCaps [188] is a large-scale dataset introduced in 2021, consisting of 12 million image-text pairs collected from Reddit. This dataset includes images and captions depicting a variety of objects and scenes, sourced from a manually curated set of subreddits to ensure diverse yet focused content. The data collection process involves three steps: subreddit selection, image post filtering, and caption cleaning. Images are primarily photographs from 350 selected subreddits, excluding any NSFW, banned, or quarantined content. Filtering techniques are used to maintain high-quality captions and mitigate privacy and harmful stereotypes, resulting in a robust and extensive dataset.

Laion-5b [189] is a large-scale vision-language dataset derived from Common Crawl, containing nearly 6 billion image-text pairs. Images with alt-text were extracted and processed to remove low-quality and malicious content. Filtering based on cosine similarity with OpenAI's ViT-B/32 CLIP model reduced the dataset size significantly. The dataset is divided into three subsets: 2.32 billion English pairs, 2.26 billion multilingual pairs, and 1.27 billion pairs with undetected languages. Metadata includes image URLs, text, dimensions, similarity scores, and NSFW tags.

DiffusionDB [190] is the first large-scale prompt dataset totaling 6.5TB, containing 14 million images generated by Stable Diffusion using 1.8 million unique prompts. Constructed by collecting images shared on the Stable Diffusion public Discord server. Most prompts are between 6 to 12 tokens long, with a significant spike at 75 tokens, indicating many users exceed the model's limit. 98.3% of the prompts are in English, with the rest covering 34 other languages. DiffusionDB provides unique research opportunities in prompt engineering, explaining large generative models, and detecting deepfakes, serving as an important resource for studying prompts in text-to-image generation and designing next-generation human-AI interaction tools.

DiffusionForensics [79] is a dataset designed for evaluating diffusion-generated image detectors. It includes 42,000 real images from LSUN-Bedroom, 50,000 from ImageNet, and 42,000 from CelebA-HQ. Generated images are produced by various models, with unconditional models like ADM, DDPM, iDDPM, and PNDM generating 42,000 images each from LSUN-Bedroom. Text-to-image models LDM, SD-v1, SD-v2, and VQ-Diffusion also generate 42,000 images each, while IF, DALLE-2, and Midjourney produce fewer images. For ImageNet, 50,000 images each are generated by a conditional model ADM and a text-to-image model SD-v1. CelebA-HQ includes 42,000 images generated by SD-v2 and smaller sets by IF, DALLE-2, and Midjourney.

LSUN Bedroom [168] dataset contains images center-cropped to 256×256 pixels. Samples are either downloaded or generated using code and pre-trained models from original publications. The dataset includes samples from ten models (e.g. ProGAN, Diff-StyleGAN2, Diff-ProjectedGAN, DDPM, IDDPM,LDM). For each model, 51,000 images were sampled, and the real part is sourced from lsun bedroom dataset [191].

DeepFakeFace (DFF) [169] is a dataset designed to evaluate deepfake detectors, featuring 120,000 images, with 30,000 real images sourced from the IMDB-WIKI dataset and 90,000 fake images. To generate these fake images, three models were used: Stable Diffusion v1.5, Stable Diffusion Inpainting, and InsightFace, each producing 30,000 images. The dataset includes high-resolution images of 512 × 512 pixels. Real images were matched by gender and age, using prompts like "name, celebrity, age" for generation. Discrepancies in facial bounding boxes were corrected using the RetinaFace detector to ensure accuracy before generating deepfakes.

RealFaces [170] consists of 25,800 images generated using Stable Diffusion, incorporating prompts for photorealistic human faces. It includes 431 images filtered by an NSFW filter, mainly depicting women and young people.

Deepart Detection Database (DDDB) [171] is designed for detecting deepfake art. It includes high-quality conventional art from LAION-5B and deepfake art from models like Stable Diffusion, DALL-E 2, Imagen, Midjourney, and Parti. Conart images are sourced from LAION-5B, while deeparts are generated using state-of-the-art models or collected from social media. DDDB consists of 64,479 conventional art images (conart) and 73,411 deepfake art images (deepart). It supports research in deepart detection, continuously updating to incorporate new deeparts and addressing privacy and storage constraints.

SynthBuster [76]. Due to the scarcity of diffusion model-generated images, SynthBuster addresses this by providing a new dataset with images from models like Stable Diffusion 1.3, 1.4, 2, and XL, Midjourney, Adobe Firefly, and DALL·E 2 and 3. While synthetic images are generated from text, SynthBuster uses the existing Raise-1k database of real images, which is a varied subset of the Raise [192] dataset, as a guideline for the generated image. Original images are not used as prompts to try to recreate or modify a similar image. They are only used as a guideline to create the new prompt for the presentation, to ensure that the resulting image is broadly in the same category as the original image. For each of the 1000 images, descriptions are generated using the Midjourney descriptor [3] and CLIP Interrogator [193]. Then, these descriptions were used as the basis for manually writing a text prompt to generate a photo-realistic image loosely based on the original image.

DE-FAKE [172] is designed for detecting AI-generated images. Real images are sourced from the MSCOCO and Flickr30k datasets. To create a corresponding set of fake images, prompts from these real images were used to generate 191,946 synthetic images through four different image generation models: Stable Diffusion, Latent Diffusion, GLIDE, and DALLE-2.

AI-Gen [173] dataset consists of 20,000 uncompressed 256 × 256 PG images from the ALASKA [194] database, which are used to construct the T2I dataset. Specific spots and objects are extracted from these Photographs (PG) images, and 5,000 prompts are generated with ChatGPT. Two AI systems, DALL·E2 [195] and DreamStudio, are used to generate four images per prompt, creating two databases: DALL·E2 [195] and DreamStudio [196]. Each database contains 20,000 Photographs (PG) images and corresponding T2I images. The images are resized to 256 × 256, 128 × 128, and 64 × 64, and JPEG compression is applied with a quality factor between 75 and 95. The datasets are divided into training (12,000 pairs), validation (3,000 pairs), and testing (5,000 pairs).

AutoSplice [175] is a image dataset containing 5,894 manipulated and authentic images, designed to aid in developing generalized detection methods. The dataset consists of 3,621 images generated by locally or globally manipulating real-world image-caption pairs from the Visual News dataset. The DALL-E2 generative model was used to create synthetic images based on text inputs. AutoSplice construction involved pre-processing with object detection and text parsing, human annotations to select and modify object descriptions, and post-processing to filter out images with visual artifacts. The final dataset includes 3,621 high-quality manipulated images and 2,273 authentic images, with versions in both lossless and gently lossy JPEG compression formats.

ArtiFact [174] is a large-scale dataset designed to evaluate the generalizability and robustness of synthetic image detectors by incorporating diverse generators, object categories, and real-world impairments. It includes 2,496,738 images, with 964,989 real and 1,531,749 fake images. The dataset covers multiple categories such as Human/Human Faces, Animal/Animal Faces, Places, Vehicles, and Art, sourced from 8 source datasets (e.g., COCO, ImageNet, AFHQ, Landscape) . It features images synthesized by 25 distinct methods, including 13 GANs (e.g., StyleGAN3, StyleGAN2, ProGAN), 7 Diffusion models (e.g., DDPM, Latent Diffusion, LaMA), and 5 other generators (e.g., CIPS, Palette). To ensure real-world applicability, images undergo impairments like random cropping, resizing, and JPEG compression according to IEEE VIP Cup 2022 standards.

CIFAKE [181] consists of 120,000 images, split evenly between real and synthetic images. The real images are taken from the CIFAR-10 [197] dataset, comprising 60,000 32x32 RGB images across ten classes: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck, with 50,000 images used for training and 10,000 for testing. The synthetic images are generated using the CompVis Stable Diffusion model (version 1.4), which is trained on subsets of the LAION-5B [189] dataset. The generation process involves reverse diffusion from noise to create 6,000 images per class, mimicking the CIFAR-10 [197] dataset. Similar to the real images, 50,000 synthetic images are used for training and 10,000 for testing, with labels indicating their synthetic nature.

GenImage [182] is designed to evaluate detectors' ability to distinguish between AI-generated and real images. It includes 2,681,167 images, with 1,331,167 real images from ImageNet and 1,350,000 fake images generated using eight models: BigGAN, GLIDE, VQDM, Stable Diffusion V1.4, Stable Diffusion V1.5, ADM, Midjourney, and Wukong. The images are balanced across ImageNet's 1000 classes, with specific allocations for training and testing. Each model generates a nearly equal number of images per class, ensuring no overlap in real images. The dataset features high variability and realism, particularly in animals and plants, providing a robust basis for developing detection models.

Fake2M [183] dataset is a large-scale collection of over 2 million AI-generated images. These images are created using three different models: Stable Diffusion v1.5, IF, and StyleGAN3. The dataset aims to investigate whether models can distinguish AI-generated images from real ones.

DiFF [180] comprises over 500,000 images synthesized using thirteen distinct generation methods under four conditions, leveraging 30,000 textual and visual prompts to ensure high fidelity and semantic consistency. The dataset includes pristine images from 1,070 celebrities, curated from sources like VoxCeleb2 and CelebA, totaling 23,661 images. Prompts, derived from these pristine images, include original and modified textual prompts as well as visual prompts. The dataset covers four categories of diffusion models: Text-to-Image (T2I), Image-to-Image (I2I), Face Swapping (FS), and Face Editing (FE), employing methods like Midjourney, Stable Diffusion XL, DreamBooth, DiffFace, and others to generate the forged images.

WildFake [184] is designed to assess the generalizability and robustness of fake image detectors. Developed with diverse content from open-source websites and generative models, it provides a comprehensive set of high-quality fake images. It includes images from DMs, GANs, and other generators, with categories such as "Early" and "Latest" models. The dataset also features nine kinds of DMs generators and various fine-tuning strategies for SD-based generators. Images were collected using a generation pipeline from platforms like Civitai and Midjourney, ensuring a representative sample of real-world quality. Real images were sourced from datasets like COCO, FFHQ, and Laion-5B. WildFake contains 3,694,313 images, with 1,013,446 real and 2,680,867 fake images, split into training and testing sets in a 4:1 ratio.

B. Video Datasets

In this section, we review key video datasets that have been pivotal in advancing state-of-the-art AI models. These resources Offer diverse video-text pairs, high-resolution clips, and specialized content, each contributing uniquely to the progress of AI-driven video technology. For a detailed comparison, refer to Table VII, which summarizes the characteristics and scope of these datasets.

YT-Tem-180M [198] was collected from 6 million public YouTube videos, totaling 180 million clips, and annotated by ASR. It includes diverse content such as instructional lifestyle vlogs, and auto-suggested videos on topics. Videos were filtered to exclude those an English ASR track, over 20 minutes long, in 'ungrounded" categories, or with thumbnails to contain objects. Each video was split into segments of an image frame and corresponding spoken words, resulting in 180 million segments.

WebVid-2M [199] is a large-scale video-text pretraining dataset consisting of 2.5 million video-text pairs. The average length of each video is 18.0 seconds, and the average caption length is 12.0 words. The raw descriptions for each video are collected from the Alt-text HTML attribute associated with web images. This dataset was scraped from the web using a method similar to Google Conceptual Captions (CC3M), which includes over 10% of images that are video thumbnails. WebVid-2M captions are manually generated, well-formed sentences aligned with the video content, contrasting with the HowTo100M [105] dataset, which contains incomplete sentences from continuous narration that may not be temporally aligned with the video.

CATER-GEN-v1 [151] is a synthetic dataset set in a 3D environment, derived from CATER [210], featuring two objects (cone and snitch) and a large table plane. It includes four atomic actions: "rotate", "contain", "pick-place", and "slide", with each video containing one or two actions. Descriptions are generated using predefined templates, with a resolution of 256x256 pixels. The dataset includes 3,500 training pairs and 1,500 testing pairs.

CATER-GEN-v2 [151] is a more complex version of CATER-GEN-v1, containing 3 to 8 objects per video, each with randomly chosen attributes from five shapes, three sizes, nine colors, and two materials. The actions are the same as in CATER-GEN-v1, but descriptions are designed to create ambiguity by omitting certain attributes. The video resolution is 256x256 pixels, and the dataset includes 24,000 training pairs and 6,000 testing pairs.

Internvid [202] is a video-centric multimodal dataset created for large-scale video-language learning, featuring high temporal dynamics, diverse semantics, and strong video-text correlations. It includes 7 million YouTube video-text correlations. It includes 7 million YouTube videos with an average duration of 6.4 minutes, covering 16 topics. Videos were collected based on popularity and action-related queries, ensuring diversity by including various countries and languages. Each video is segmented into clips, resulting in 234 million clips from 2s to more than 30s duration, which were captioned using a multiscale method focusing on common objects and actions. InternVid emphasizes high resolution, with 85% of videos at 720P, and provides comprehensive multimodal data including audio, metadata, and subtitles. The dataset is notable for its action-oriented content, containing significantly more verbs compared to other datasets, and includes 7.1 million interleaved video-text data pairs for in-context learning.

FlintstonesHD [153] is a densely annotated long video dataset created to promote the development of long video generation. The dataset is built from the original Flintstones cartoon, containing 166 episodes with an average of 38,000 frames per episode, each at a resolution of 1440 × 1080 pixels. Unlike existing video datasets, FlintstonesHD addresses issues such as short video lengths, low

TABLE VII. Video Datasets. Datasets With Grey Background Are Used in A AI-generated Videos Detection

Dataset	Year	Source	Size	Domain	Resolution	Text	Avg len (sec)	Duration (hrs)	Unique Features
YT-Tem-180M [198]	2021	YouTube, HowTo100M	180M Videos 180M Text	Open	-	ASR	-	-	Filters to exclude non-English ASR and visuall «ungrounded» categories
WebVid-2M [199]	2021	Web	2.5M Videos 2.5M Text	Open	360p	Manual	18.0	13K	Manually generated captions, aligned with video content
WebVid-10M [199]	2021	Web	10M Videos 10M Text	Open	360p	Alt-Text	18.0	52K	Manually generated captions, aligned with video content
CATER-GEN-v1 [151]	2022	Synthetic 3D objects	5K Video 5K Text	Geometric	256p	Predefined template	-	-	Synthetic, simple scenes with atomic actions
CATER-GEN-v2 [151]	2022	Synthetic 3D objects	30K Video 30K Text	Geometric	256p	Predefined Template	-	-	Increased complexity with more objects and attributes
CelebV-HQ [200]	2022	Web	35,666 Videos	Face	512p	Manual	3 to 20	65	High-quality, detailed text descriptions
HD-VILA-100M [201]	2022	YouTube	103M Videos 103M Text	Open	720p	ASR	13.4	371.5K	High-quality alignment of videos and transcriptions
Internvid [202]	2023	YouTube	7.1M Videos 234M clips	Open	360p 512p 720p	Generated	11.7	760.3K	Action-oriented, diverse languages, and high video-text correlation
FlintstonesHD [153]	2023	Flintstones cartoon	166 episodes	Cartoon	1440x1080	Generated	-	-	Densely annotated for long video generation
Celebv-text [203]	2023	Web	70K Videos 1.4M Text	Face	512p+	Semi-Auto Generated	<5s	279	High-quality, detailed text descriptions
HD-VG-130M [204]	2023	YouTube	130M Videos 130M Text	Open	720p	Generated	~ 5.1	184K	High-definition, single-scene clips
Youku-mPLUG [205]	2023	Youku platform	10M Videos 10M Text	Open	-	-	54.2	150K	Focused on advancing Chinese multimodal LLMs
VidProM [206]	2024	Pika Discord	1.67M prompts 6.69M Videos	Open	-	Manual		-	Extensive prompts with semantic uniqueness
MiraData [207]	2024	YouTube, Videvo, Pixabay, Pexels HD-VILA-100M		Open	720p	Generated	72.1	16K	High visual quality, detailed captions
GenVideo [162]	2024	Kinetics-400 Youku-mPLUG MSR-VTT Video Gen Methods	~ 2.31M Videos	Open	-	Automatic	2 to 6		Balance of real and fake videos across diverse scenes
ExposingAI-Video [160]	2024	MSVD, Potat1 Ali-vilab,ZScope T2V-zero	2K Videos	Open	-	Automatic	-	-	H. 265 compression and quality degradation simulation
Synth-vid-detect [159]	2024	MIT, Video-ACID Gen Video Methods	18.75K Videos	Open	-	Automatic	-	-	H. 265 compression Out-of-distribution test set
GVD [163]	2024	GOT, Youtube_vos2 Gen Video Methods	-	Open	-	Automatic	-	-	Collection from various SOTA models
GVF [164]	2024	MSVD, MSR-VTT Gen Video Methods	964 Videos 964 Text	Open	-	Automatic	-	-	Diversity in forgery targets, scenes, and behaviors
GenVidDet [165]	2024	InternVid, HD-VG-130M Gen Video Methods	~2.66M Videos	Open	256p 512p 720p	Automatic	-	4442	Large-scale dataset cover diverse content
TOINR [167]	2024	VidVRD, SVD-XT YouTube SORA, Pika, GEN-2	~2.826K Videos	Open	-	Automatic	-	-	Out-domain testing with various generation tools
Panda-70m [208]	2024	HD-VILA-100M	70.8M Videos 70.8M Text	Open	720p	Automatic	8.5s	166.8K	High-quality captions with significant improvements in downstream tasks
VAST-27M [209]	2024	HD-VILA-100M	27M Videos 297M Text	Open	-	Generated	5 to 30 sec	-	Comprehensive with vision, audio, and omni-modality captions

resolution, and coarse annotations. The image captioning model GIT2 [211] was used to generate dense captions for each frame, with manual filtering to correct errors, thus providing detailed annotations that capture movement and story nuances. This dataset serves as a benchmark for improving long video generation.

Celebv-text [203] is a large-scale facial text-video dataset aimed at providing high-quality video samples with relevant, diverse text descriptions. Constructed through data collection and processing, data annotation, and semi-auto text generation, it features 70, 000 video clips totaling around 279 hours. Videos were sourced from the internet, using queries like human names and movie titles, excluding low-resolution and short clips, and processed to maintain high quality without upsampling or downsampling. Annotations include static attributes like general appearance and light conditions, and dynamic attributes like actions and emotions, with both automatic and manual methods used for accuracy. Texts were generated using a combination of manual descriptions and auto-generated templates based on common grammar structures, resulting in longer and more detailed text descriptions compared to other datasets. CelebV-Text surpasses existing datasets like MM-Vox [212] and CelebV-HQ [200] in scale, resolution, and text-video relevance, offering a comprehensive resource for facial video analysis.

VidProM [206] is a large-scale dataset for text-to-video diffusion models, collected from Pika Discord channels between July 2023 and February 2024.It includes 1,672,243 unique text-to-video prompts, embedded with 3072-dimensional embeddings using OpenAI's text-embedding-3-large API. The dataset includes NSFW probabilities assigned using the Detoxify model, with less than 0.5% of prompts flagged as potentially unsafe. It features 6.69 million videos generated by Pika, VideoCraft2, Text2Video-Zero, and ModelScope, involving significant computational resources. After filtering for semantic uniqueness, VidProM retains 1,038,805 unique prompts. Compared to DiffusionDB, VidProM has 40.6% more semantically unique prompts and supports longer, more complex prompts due to its advanced embedding model. VidProM includes videos generated by four state-of-the-art models, resulting in over 14 million seconds of video content. VidProM's extensive video content and complex prompts, requiring dynamic and temporal descriptions, make it a valuable resource for developing text-to-video generative models.

MiraData [207] is a large-scale text-video dataset with long durations and detailed structured captions. The dataset, finalized through a five-step process, sources videos from YouTube, Videvo, Pixabay, and Pexels to ensure diverse content and high visual quality. From YouTube, 156 high-quality channels were selected, resulting in 68K videos and 173K clips post-processing. Additional videos were sourced from HD-VILA-100M, Videvo (63K), Pixabay (43K), and Pexels (318K). Video clips were split and stitched using models like Qwen-VL-Chat and DINOv2, ensuring semantic coherence and content continuity. MiraData provides five versions of filtered data based on video color, aesthetic quality, motion strength, and NSFW content, with 788K to 9K clips. Captions were generated using GPT-4V, resulting in dense and structured descriptions with average lengths of 90 and 214 words respectively. MiraData surpasses previous datasets in visual quality and motion strength, making it ideal for text-to-video generation tasks.

GenVideo [162] is a large-scale dataset developed to evaluate the generalizability and robustness of AI-generated video detection models. The training set contains 2, 294, 594 video clips, including 1, 213, 511 real and 1, 081, 083 fake videos, while the testing set includes 19, 588 video clips, with 10, 000 real and 8, 588 fake videos. The dataset features high-quality fake videos sourced from open-source websites and various pre-trained models, covering a wide range of scenes such as landscapes, people, buildings, and objects. Video duration's range from 2 to 6 seconds, with diverse aspect ratios. Real videos are sourced from datasets like Kinetics-400, Youku-mPLUG, and MSR-VTT [213]. Fake videos are generated using diffusion-based models, auto-regressive models, and other methods such as VideoPoet, Emu, Sora, VideoCrafter, latent flow diffusion models, masked generative video transformer, and autoregressive models. Additionally, sources include external web scraping and service-based methods like the Pika website. This diverse and comprehensive collection aims to enhance the understanding and detection of AI-generated videos across numerous real-world contexts.

ExposingAI-Video [160] is composed of 1,000 natural videos sourced from the MSVD [214] dataset, paired with 1,000 fake videos generated using four advanced diffusion-based video generators, resulting in 96,000 fake frames. The dataset offers diverse content driven by text prompts, featuring rich motion information distinct from static images. It includes videos generated by models such as ali-vilab, zeroscope, potat1, and a zero-shot text-to-video model, each providing unique configurations. Additionally, the dataset incorporates three video post-processing operations—H.265 ABR compression, H.265 CRF compression, and Bit Error—to simulate quality degradation for robustness evaluation.

Synth-vid-detect [159] consists of both real and synthetic videos for training and evaluation. It includes 7,654 real videos for training, 784 for validation, and 1,661 for testing, sourced from the Moments in Time (MIT) [215] and Video-ACID [216] datasets. The synthetic videos, totaling 6,197 for training, 624 for validation, and 1,429 for testing, were generated using Luma, VideoCrafter-v1, CogVideo, and Stable Video Diffusion, with diverse scenes and activities represented. All videos were compressed using H.264 at a constant rate factor of 23. For testing, an exclusive set of prompts and videos was used to avoid overlap with the training data. Additionally, the dataset includes an out-of-distribution, test-only set of 401 synthetic videos generated by Sora, Pika, and VideoCrafter-v2.

Generated Video Dataset (GVD) [163] includes 11,618 video samples produced by 11 different state-of-the-art generator models. These models generate videos using either T2V or I2V techniques. The dataset was primarily collected from the Discord platform, where users share videos generated by various models. For training and validation, 550 T2V-generated videos from Moonvalley [217] and 550 real videos from the YouTube_vos2 [218] dataset were used. All generated videos not used in training and validation are designated for testing, with real test videos sourced from the GOT [219] dataset.

GeneratedVideoForensics (GVF) [164] dataset consists of 964 triples, each containing a real video, a corresponding text prompt, and a video generated by one of four different open-source text-to-video generation models: Text2Video-zero, ModelScopeT2V, ZeroScope, and Show-1. These models cover various forgery targets, scenes, behaviors, and actions, ensuring the dataset's diversity. The real videos and prompts were collected from MSVD [214] and MSR-VTT [213] datasets, with a focus on simulating realistic video distributions across spatial and temporal dimensions. It also includes vidoes from most popular commercial models like OpenAI's Sora, Pika, Gen-2 and Google's Veo.

GenVidDet [165] is a large-scale video dataset created for AI-generated video detection, comprising over 2.66 million clips with more than 4442 hours of content. It includes real videos sourced from the InternVid [202] and HD-VG-130M [204] datasets, totaling over 1.46 million clips, and AI-generated videos from the VidProM dataset using four different models, adding approximately 1.12 million clips. Additionally, new AI-generated videos were created using the latest models like Open-Sora, StreamingT2V and DynamiCrafter to enhance the dataset's diversity.

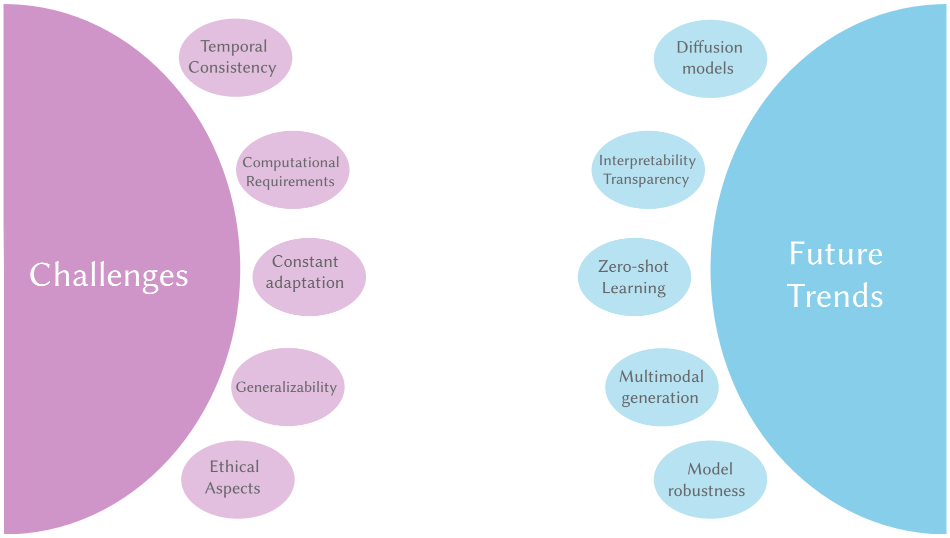


Fig. 11. Overview of trends and challenges in the generation and detection of AI-generated image and video samples.

Turns Out I'm Not Real (TOINR) [167] dataset was constructed to evaluate a method using public video generation tools, including Stable Video Diffusion (SVD), Pika, Gen-2, and SORA. The dataset includes 1,000 real video clips from the ImageNet Video Visual Relation Detection (VidVRD) [220] dataset and 1,000 fake video clips generated with SVD-XT [89]. It also comprises an additional real and fake clips for out-domain testing: 107 real (VidVRD) and 107 fake clips generated with Pika, 107 (VidVRD) real and 107 fake clips generated with Gen-2, and 207 real and 191 fake clips sourced from YouTube and SORA website.

HD-VILA-100M [201] is a high-resolution and diversified video-language dataset designed to overcome limitations in existing datasets. Introduced to aid tasks such as Text-to-video retrieval and video QA, it comprises 103 million video clip and sentence pairs from 3.3 million videos, totaling 371.5K hours. Sourced from diverse YouTube content, including professional channels like BBC Earth and National Geography, HD-VILA-100M emphasizes quality and alignment of videos and transcriptions. Only videos with subtitles and 720p resolution were included, resulting in a final set of 3.3 million videos, balanced across 15 categories. For video-text pairing, the dataset utilizes video transcriptions instead of manual annotations, offering richer information. Subtitles, often generated by ASR, were split into complete sentences using an off-the-shelf tool. Sentences were aligned with video clips using Dynamic Time Warping, producing pairs averaging 13.4 seconds in length and 32.5 words per sentence.

HD-VG-130M [204] is a large-scale dataset for Text-to-video generation, comprising 130 million text-video pairs from the open domain. Created to address limitations in existing datasets, it features high-definition (720p), widescreen, and watermark-free videos. Collected from YouTube, the videos were processed using PySceneDetect for scene detection, resulting in single-scene clips of less than 20 seconds each. Captions were generated using BLIP-2, ensuring that descriptions, typically around 10 words, are representative of the visual content. Covering 15 categories, HD-VG-130M provides diverse and high-quality data for training video generation models.

Youku-mPLUG [205] is the first Chinese video-language pretraining dataset, released in 2023 and collected from the Youku video-sharing platform. It comprises 10 million high-quality Chinese video-text pairs filtered from 400 million raw videos, covering 45 diverse categories with an average video length of 54.2 seconds. This dataset was created to advance Vision-language pre-training (VLP) and multimodal Large language models (LLMs) within the Chinese community. Strict criteria for safety, diversity, and quality were applied, involving multi-level risk detection to eliminate high-risk content and video fingerprinting to ensure a balanced distribution. Additionally, the dataset includes 0.3 million videos for downstream benchmarks, designed to assess video-text retrieval, video captioning, and video category classification tasks.

Panda-70m [208] is a large-scale video dataset created for video captioning, video and text retrieval, and text-driven video generation. It consists of 70 million high-resolution, semantically coherent video clips with captions. The dataset was developed from 3.8 million long videos collected from HD-VILA-100M [201]. To generate accurate captions, a two-stage semantics-aware splitting algorithm was used, followed by multiple cross-modality teacher models to predict candidate captions. A subset of 100,000 videos was manually annotated to fine-tune a retrieval model, which then selected the best captions for the entire dataset. Panda-70M addresses the challenge of collecting high-quality video-text data and shows significant improvements in downstream tasks. The dataset primarily contains vocal-intensive videos such as news, TV shows, and documentaries.

VAST-27M [209] consists of a total of 27 million video clips covering diverse categories, each paired with 11 captions (5 vision, 5 audio, and 1 omni-modality). The average lengths of vision, audio, and omni-modality captions are 12.5, 7.2, and 32.4 words respectively. The dataset bridges various modalities including vision, audio, and subtitles in videos. The clips were selected from the HD-VILA-100M dataset [201], ensuring each clip is between 5 and 30 seconds long and contains all three modalities. Vision captions were generated using a model trained on corpora such as MSCOCO, VATEX, MSRVTT, and MSVD [214], while audio captions were generated using VALOR-1M and WavCaps datasets. An LLM, Vicuna-13b, was used to integrate these captions into a single omni-modality caption. VAST-27M spans over 15 categories, including music, gaming, education, entertainment, and animals. its comprehensiveness, the dataset may inherit biases from the corpora and models used in its creation, highlighting the need for more diverse and larger-scale omni-modality corpora.

VI. Challenges and Future Trends

Throughout this state-of-the-art review we have analysed the most recent approaches and methodologies for the generation and detection of synthetic video and image samples. This has given us a global view of the area, as well as a glimpse of current research trends and the challenges researchers will have to face in the coming years, see Fig. 11.

First of all, we will focus on analysing the trends that will drive research in the area in the coming years, based on the results obtained from this analysis.

1. **Sample generation with diffusion models**, where the diffusion process in these models involves iterating over the input data and gradually refining the generation to fit a target distribution or to achieve the desired effect. As we have been able to observe throughout the different sections related to the generation of samples, whether video or image, the diffusion models seem to be predominating over the rest of the generation techniques, such as autoencoders or GANs. Taking into account all the research being carried out in this domain, it would not be surprising to see it monopolises multimedia content generation techniques in the coming years.
2. **Zero-Shot Learning**. This learning approach is a game changer, as it allows generative models to create content in new domains, even with entirely new features, without needing to be trained with data from those exact situations. This makes it possible, within generative techniques, to generate a wide range of content, even when a large amount of labelled data is not available. But it remains difficult to develop models capable of accurately understanding and generating content in completely new contexts. Regarding detection, zero-shot learning has the potential to help identify AI-generated content in many different data types and formats, even in the absence of huge curated datasets. However, the wide variety of synthetic content creation methods makes it difficult to create perfectly adapted detection models. Further research is needed to determine how to improve the generalisability of these models.
3. **Interpretability and Transparency**. As the content generated by AI becomes more sophisticated, it becomes increasingly important to ensure that detection models are not only effective, but also easy to understand. Users need to be convinced that the model is making the right decisions, which means that the model needs to provide clear and understandable reasons for why it has identified something as synthetic. In addition, these techniques allow us to understand whether the features that the models are using to achieve at the output are adequate or whether the system has deficiencies or biases. Therefore, the application of explainability techniques has many advantages.
4. **Multimodal data generation**. As we have seen in Section V, multimodal sample generation techniques are the least explored of all. The main reason may be their complexity, as a very precise synchronisation between video and audio has to be achieved. However, it is quite possible that this approach will start to become more relevant, due to the opportunities it presents. Regarding synthetic multimodal data detection techniques, research will be extremely limited until quality datasets are available to train robust models, capable of being applied to real situations.
5. **Model robustness**. Detection models must be able to robustly withstand various transformations and adversarial attacks, such as image compression, blurring or text paraphrasing, which can significantly degrade detection performance. The ability to withstand such manipulations is crucial for the reliable identification of synthetic content in various real-world scenarios. These types of distortions can effectively compromise a model's ability to correctly identify synthetic content. So being able to overcome these challenges is essential to ensure that the model works reliably in all kinds of scenarios.

Finally, we are going explore the different ***challenges*** that the field of video and image generation is likely to face. This review has highlighted several weaknesses that must be addressed, as they represents significant obstacles for future research in this domain.

1. **Temporal Consistency**. One of the main problems in the generation of synthetic video samples is the formation of artefacts or inconsistencies between the created frames. Smooth and realistic motion patterns are essential for video sequences, however generative models may find it difficult to maintain this from frame to frame. In addition, inconsistent frame transitions can lead to visual artifacts such as flicker, which affect the realism of the generated content. Although advances in techniques such as Implicit Neural Representations (INR), interplacing of multiple temporal attention layers, fully fine tuning on video datasets, as well as hierarchical discriminators have shown promise, further research is necessary to achieve smooth and realistic video sequences.
2. **Computational Requirements**. Video generation and detection involves processing high dimensional data, which significantly increases the computational requirements for training and inference, which can be an obstacle for small organizations. Developing more efficient algorithms and parallelization techniques for video generation is an ongoing challenge.
3. **Constant adaptation**: as we have seen in this survey, there are two main lines of research: the generation of synthetic samples and their detection techniques. Every day there are new, more sophisticated generation techniques that generate more realistic samples, so new detection models that are capable of distinguishing these synthetic samples from the real ones have to be constantly developed, i.e. it is a race. As well as the development of new quality datasets that will be the starting point of the detection systems. Another approach may be the periodic retraining of models. Whether to simply re-train a model from scratch or continue to update it through continuous learning is an ongoing challenge that researchers are still working on.
4. **Generalizability of Detection Models.** A key challenge for detection models is to be able to handle new data and new models. Generative AI models (GAIMs) evolve rapidly and if a detection model is too focused on the specific data it has been trained on, it tends to struggle with new, unseen data and updated models. To remain relevant and effective, detection models must be able to generalise to different datasets and types of generative architectures.
5. **Ethical Aspects**. The realistic nature of AI-generated content raises serious ethical questions, particularly when it comes to potential misuse. Deepfakes, fake news and other misleading content can cause real harm. To combat this, it is not enough to develop effective detection methods. We also need ethical guidelines, regulations and access controls to prevent AI technology from being used in harmful ways.

VII. Conclusions

Generative AI has witnessed exponential growth in recent years, exemplified by tools like ChatGPT that showcase its advancing capabilities. Multimedia content generation models have achieved remarkable performance across a variety of tasks, offering substantial benefits to domains such as entertainment, education, and cybersecurity. However, these advancements also introduce risks that cannot be ignored. Alongside the development of new generative AI models for producing high-quality multimedia content, there is a critical need to create detection systems that can be effectively applied in real-world situations.

This review aims to address these dual objectives by providing a comprehensive analysis of synthetic image and video generation techniques, as well as the methods used for their detection. It also examines the principal datasets available in the current state of the art and explores future trends and challenges faced by researchers in the field. By critically evaluating the existing technologies for generating and detecting multimedia content, we seek to define the research directions that should be pursued in the coming years. The insights gathered from this survey are intended to facilitate and stimulate further research on generative AI techniques for multimedia content, ultimately contributing to both the advancement of the field and the mitigation of associated risks.

Acknowledgment

This work has been partially supported by the project PCI2022-134990-2 (MARTINI) of the CHISTERA IV Cofund 2021 program; by MCIN/AEI/10.13039/501100011033/ and European Union NextGenerationEU/PRTR for XAI-Disinfodemics (PLEC 2021-007681) grant, by European Comission under IBERIFIER Plus - Iberian Digital Media Observatory (DIGITAL-2023-DEPLOY- 04-EDMO-HUBS 101158511), and by TUAI Project (HORIZON-MSCA-2023-DN-01-01, Proposal number: 101168344); by EMIF managed by the Calouste Gulbenkian Foundation, in the project MuseAI; and by Comunidad Autonoma de Madrid, CIRMA-CM Project (TEC-2024/COM-404). Abdenour Hadid is funded by TotalEnergies collaboration agreement with Sorbonne University Abu Dhabi.

References

[1] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.

[2] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," 2022. [Online]. Available: https://arxiv.org/abs/2112.10741.

[3] Midjourney, "Midjourney platform." Online. [Online]. Available: https://www.midjourney.com/home, Accessed: Nov. 07, 2024.

[4] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in neural information processing systems*, vol. 35, pp. 36479–36494, 2022.

[5] D. Kondratyuk, L. Yu, X. Gu, J. Lezama, J. Huang, R. Hornung, H. Adam, H. Akbari, Y. Alon, V. Birodkar, *et al.*, "Videopoet: A large language model for zero-shot video generation," *arXiv preprint arXiv:2312.14125*, 2023.

[6] OpenAI, "Sora: Video generation models as world simulators," OpenAI, 2024. [Online]. Available: https://openai.com/index/sora/, Accessed: Nov. 07, 2024.

[7] J. Bruce, M. D. Dennis, A. Edwards, J. Parker- Holder, Y. Shi, E. Hughes, M. Lai, A. Mavalankar, R. Steigerwald, C. Apps, *et al.*, "Genie: Generative interactive environments," in *Proceedings of the 41st International Conference on Machine Learning*, vol. 235 of *Proceedings of Machine Learning Research*, 21–27 Jul 2024, pp. 4603–4623, PMLR.

[8] G. Madaan, S. K. Asthana, J. Kaur, "Generative ai: Applications, models, challenges, opportunities, and future directions," *Generative AI and Implications for Ethics, Security, and Data Management*, pp. 88–121, 2024.

[9] X. Zhao, X. Zhao, "Application of generative artificial intelligence in film image production," *Computer- Aided Design & Applications*, vol. 21, pp. 29–43, 2024, doi: 10.14733/cadaps.2024.S27.29-43.

[10] Á. Huertas-García, H. Liz, G. Villar-Rodríguez, Martín, J. Huertas-Tato, D. Camacho, "Aida- upm at semeval-2022 task 5: Exploring multimodal late information fusion for multimedia automatic misogyny identification," in *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, 2022, pp. 771–779.

[11] N. Anantrasirichai, D. Bull, "Artificial intelligence in the creative industries: a review," *Artificial intelligence review*, vol. 55, no. 1, pp. 589–656, 2022.

[12] H. Choi, *Generative AI Art Exploration and Image Generation Fine Tuning Techniques*. PhD dissertation, California Institute of the Arts.

[13] A. Doe, B. Smith, C. White, "Gans for medical image synthesis: A comprehensive review," *Medical Image Analysis*, vol. 78, p. 102345, 2023.

[14] U. Mittal, S. Sai, V. Chamola, *et al.*, "A comprehensive review on generative ai for education," *IEEE Access*, vol. 12, pp. 142733–142759, 2024.

[15] H. S. Mavikumbure, V. Cobilean, C. S. Wickramasinghe, D. Drake, M. Manic, "Generative ai in cyber security of cyber physical systems: Benefits and threats," in *2024 16th International Conference on Human System Interaction (HSI)*, 2024, pp. 1–8, IEEE.

[16] S. Oh, T. Shon, "Cybersecurity issues in generative ai," in *2023 International Conference on Platform Technology and Service (PlatCon)*, 2023, pp. 97–100, IEEE.

[17] H. Liz-Lopez, M. Keita, A. Taleb-Ahmed, A. Hadid, J. Huertas-Tato, D. Camacho, "Generation and detection of manipulated multimodal audiovisual content: Advances, trends and open challenges," *Information Fusion*, vol. 103, p. 102103, 2024.

[18] A. Giron, J. Huertas-Tato, D. Camacho, "Multimodal analysis for identifying misinformation in social networks," in *The 2024 World Congress on Information Technology Applications and Services*, 2024, World IT Congress 2024.

[19] K. Shiohara, T. Yamasaki, "Detecting deepfakes with self-blended images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18720–18729.

[20] A. Martín, A. Hernández, M. Alazab, J. Jung, D. Camacho, "Evolving generative adversarial networks to improve image steganography," *Expert Systems with Applications*, vol. 222, p. 119841, 2023.

[21] Á. Huertas-García, A. Martín, J. Huertas-Tato, D. Camacho, "Camouflage is all you need: Evaluating and enhancing transformer models robustness against camouflage adversarial attacks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2024.

[22] T. Zhang, "Deepfake generation and detection, a survey," *Multimedia Tools and Applications*, vol. 81, no. 5, pp. 6259–6276, 2022.

[23] S. Tyagi, D. Yadav, "A detailed analysis of image and video forgery detection techniques," *The Visual Computer*, vol. 39, no. 3, pp. 813–833, 2023.

[24] Z. Jia, Z. Zhang, L. Wang, T. Tan, "Human image generation: A comprehensive survey," *ACM Computing Surveys*, 2022.

[25] A. Figueira, B. Vaz, "Survey on synthetic data generation, evaluation methods and gans," *Mathematics*, vol. 10, no. 15, p. 2733, 2022.

[26] T. T. Nguyen, Q. V. H. Nguyen, D. T. Nguyen, D. T. Nguyen, T. Huynh-The, S. Nahavandi, T. T. Nguyen, Q.-V. Pham, C. M. Nguyen, "Deep learning for deepfakes creation and detection: A survey," *Computer Vision and Image Understanding*, vol. 223, p. 103525, 2022.

[27] A. Bauer, S. Trapp, M. Stenger, R. Leppich, S. Kounev, M. Leznik, K. Chard, I. Foster, "Comprehensive exploration of synthetic data generation: A survey," *arXiv preprint arXiv:2401.02524*, 2024.

[28] P. Cao, F. Zhou, Q. Song, L. Yang, "Controllable generation with text-to-image diffusion models: A survey," *arXiv preprint arXiv:2403.04279*, 2024.

[29] I. Joshi, M. Grimmer, C. Rathgeb, C. Busch, F. Bremond, A. Dantcheva, "Synthetic data in human analysis: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 7, pp. 4957–4976, 2024, doi: 10.1109/TPAMI.2024.3362821.

[30] P. Cao, F. Zhou, Q. Song, L. Yang, "Controllable generation with text-to-image diffusion models: A survey," 2024. [Online]. Available: https://arxiv.org/abs/2403.04279.

[31] T. Zhang, Z. Wang, J. Huang, M. M. Tasnim, W. Shi, "A survey of diffusion based image generation models: Issues and their solutions," 2023. [Online]. Available: https://arxiv.org/abs/2308.13142.

[32] A. Sauer, T. Karras, S. Laine, A. Geiger, T. Aila, "Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis," in *International conference on machine learning*, 2023, pp. 30105–30118, PMLR.

[33] M. Kang, J.-Y. Zhu, R. Zhang, J. Park, E. Shechtman, S. Paris, T. Park, "Scaling up gans for text-to-image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10124–10134.

[34] H. Ku, M. Lee, "Textcontrolgan: Text-to-image synthesis with controllable generative adversarial networks," *Applied Sciences*, vol. 13, no. 8, p. 5098, 2023.

[35] M. Tao, B.-K. Bao, H. Tang, C. Xu, "Galip: Generative adversarial clips for text-to-image synthesis," in *Proceedings of the IEEE/CVF Conference on*

Computer Vision and Pattern Recognition, 2023, pp. 14214–14223.

[36] Y. A. Ahmed, A. Mittal, “Unsupervised co-generation of foreground-background segmentation from text- to-image synthesis,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, vol. 12, 2024, pp. 5058–5069.

[37] Y. Xu, Y. Zhao, Z. Xiao, T. Hou, “Ufogen: You forward once large scale text-to-image generation via diffusion gans,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 8196–8206.

[38] J. Shi, C. Wu, J. Liang, X. Liu, N. Duan, “Divae: Photorealistic images synthesis with denoising diffusion decoder,” *arXiv preprint arXiv:2206.00386*, 2022.

[39] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*, 2021, pp. 8748–8763, PMLR.

[40] H. Chang, H. Zhang, J. Barber, A. Maschinot, J. Lezama, L. Jiang, M.-H. Yang, K. Murphy, W. T. Freeman, M. Rubinstein, Y. Li, D. Krishnan, “Muse: Text-to-image generation via masked generative transformers,” 2023. [Online]. Available: https://arxiv.org/abs/2301.00704.

[41] M. Ding, Z. Yang, W. Hong, W. Zheng, C. Zhou, D. Yin, J. Lin, X. Zou, Z. Shao, H. Yang, *et al.*, “Cogview: Mastering text-to-image generation via transformers,” *Advances in neural information processing systems*, vol. 34, pp. 19822–19835, 2021.

[42] M. Ding, W. Zheng, W. Hong, J. Tang, “Cogview2: Faster and better text-to-image generation via hierarchical transformers,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 16890– 16902, 2022.

[43] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, Ommer, “High-resolution image synthesis with latent diffusion models,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10684–10695.

[44] A. Razzhigaev, A. Shakhmatov, A. Maltseva, Arkhipkin, I. Pavlov, I. Ryabov, A. Kuts, Panchenko, A. Kuznetsov, D. Dimitrov, “Kandinsky: an improved text-to-image synthesis with image prior and latent diffusion,” *arXiv preprint arXiv:2310.03502*, 2023.

[45] J. Yang, J. Feng, H. Huang, “Emogen: Emotional image content generation with text-to-image diffusion models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 6358–6368.

[46] H. Li, C. Shen, P. Torr, V. Tresp, J. Gu, “Self- discovering interpretable diffusion latent directions for responsible text-to-image generation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 12006–12016.

[47] J. Ho, T. Salimans, “Classifier-free diffusion guidance,” 2022. [Online]. Available: https://arxiv.org/abs/2207.12598.

[48] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, R. Rombach, “Sdxl: Improving latent diffusion models for high-resolution image synthesis,” *arXiv preprint arXiv:2307.01952*, 2023.

[49] Z. Xue, G. Song, Q. Guo, B. Liu, Z. Zong, Y. Liu, P. Luo, “Raphael: Text-to-image generation via large mixture of diffusion paths,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[50] G. DeepMind, “Imagen 2.” http://tinyurl.com/3pakj3mk, 2023.

[51] J. Betker, G. Goh, L. Jing, T. Brooks, J. Wang, L. Li, L. Ouyang, J. Zhuang, J. Lee, Y. Guo, *et al.*, “Improving image generation with better captions,” *Computer Science. https://cdn.openai.com/papers/dall-e- 3.pdf*, vol. 2, no. 3, p. 8, 2023.

[52] L. Chen, W. Zhao, L. Xu, “Augmented cyclegan for enhanced image-to-image translation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2345–2354.

[53] Y. Wang, K. Liu, H. Zhang, “Dualgan++: Robust and efficient image-to-image translation,” *IEEE Transactions on Image Processing*, vol. 32, pp. 678–690, 2023.

[54] M. Li, E. Johnson, R. Wang, “Cut++: Enhanced contrastive unpaired translation for image synthesis,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2023, pp. 3456–3465.

[55] T. Nguyen, W. Huang, S. Lee, “Spade++: Spatially- adaptive gans for high-resolution image synthesis,” *Pattern Recognition*, vol. 122, pp. 108–119, 2022.

[56] S. Kim, D. Park, M. Lee, “Self-supervised image translation gan for high-quality synthetic image generation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4567–4576.

[57] H. Zhang, Y. Wang, K. Liu, “Unified multimodal gan for diverse image-to-image translation,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, pp. 234–245, 2024.

[58] M. Lee, S. Kim, D. Park, “Zero-shot gans: Generating images without extensive labeled data,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, pp. 567–578, 2024.

[59] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, T. Aila, “Alias-free generative adversarial networks,” in *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[60] T. Karras, T. Aila, S. Laine, J. Lehtinen, “Progressive growing of gans for improved quality, stability, and variation,” *arXiv preprint arXiv:1710.10196*, 2017.

[61] J. Smith, J. Doe, A. Brown, “Efficientgan: Reducing the computational cost of gans while preserving image quality,” *Journal of Machine Learning Research*, vol. 23, pp. 1234–1256, 2022.

[62] E. Johnson, R. Wang, M. Li, “High-resolution image synthesis with latent diffusion models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1204–1213.

[63] D. Torbunov, Y. Huang, H. Yu, J. Huang, S. Yoo, M. Lin, B. Viren, Y. Ren, “Uvcgan: Unet vision transformer cycle-consistent gan for unpaired image- to-image translation,” in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2023, pp. 702–712.

[64] W. Harvey, S. Naderiparizi, F. Wood, “Conditional image generation by conditioning variational auto- encoders,” *arXiv preprint arXiv:2102.12037*, 2022.

[65] A. Razavi, A. van den Oord, O. Vinyals, “Hierarchical variational autoencoders for high-resolution image synthesis,” *Nature*, vol. 570, pp. 234–239, 2022.

[66] A. Vahdat, J. Kautz, “Nvae: A deep hierarchical variational autoencoder,” *arXiv preprint arXiv:2007.03898*, 2022.

[67] J.-Y. Zhu, T. Park, A. A. Efros, “Stylevae: Variational autoencoders with style transfer for image synthesis,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, pp. 2345–2356, 2023.

[68] H. Kim, A. Mnih, “Factorized hierarchical variational autoencoders for disentangled representation learning,” *Journal of Machine Learning Research*, vol. 24, pp. 3456–3465, 2023.

[69] D. E. Diamantis, P. Gatoula, D. K. Iakovidis, “Endovae: Generating endoscopic images with a variational autoencoder,” in *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, 2022, pp. 1–5, IEEE.

[70] R. Dos Santos, J. Aguilar, “A synthetic data generation system based on the variational-autoencoder technique and the linked data paradigm,” *Progress in Artificial Intelligence*, pp. 1–15, 2024.

[71] S. An, J.-J. Jeon, “Distributional learning of variational autoencoder: Application to synthetic data generation,” in *Advances in Neural Information Processing Systems*, vol. 36, 2023, pp. 57825–57851, Curran Associates, Inc.

[72] G. Parmar, K. Kumar Singh, R. Zhang, Y. Li, J. Lu, J.-Y. Zhu, “Zero-shot image-to-image translation,” in *ACM SIGGRAPH 2023 Conference Proceedings*, 2023, pp. 1–11.

[73] A. Brock, J. Donahue, K. Simonyan, “Large scale gan training for high fidelity natural image synthesis,” 2019. [Online]. Available: https://arxiv.org/abs/1809.11096.

[74] S. Sinitsa, O. Fried, “Deep image fingerprint: Towards low budget synthetic image detection and model lineage analysis,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 4067–4076.

[75] N. Poredi, D. Nagothu, Y. Chen, “Ausome: authenticating social media images using frequency analysis,” in *Disruptive Technologies in Information Sciences VII*, vol. 12542, 2023, pp. 44–56, SPIE.

[76] Q. Bammey, “Synthbuster: Towards detection of diffusion model generated images,” *IEEE Open Journal of Signal Processing*, vol. 5, pp. 1–9, 2023, doi: 10.1109/OJSP.2023.3337714.

[77] T. Alzantot, C. Shou, M. Farag, Z. J. Wang, S. Pandey, M. Esmaili, “Wavelet-packets for deepfake image analysis and detection,” *Machine Learning*, vol. 111, no. 11, pp. 1–25, 2022, doi: 10.1007/s10994-022-06225-5.

[78] N. Zhong, Y. Xu, Z. Qian, X. Zhang, “Rich and poor texture contrast: A simple yet effective approach for ai-generated image detection,” *arXiv*

preprint arXiv:2311.12397, 2023.

[79] Z. Wang, J. Bao, W. Zhou, W. Wang, H. Hu, H. Chen, H. Li, "Dire for diffusion-generated image detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 22445–22455.

[80] R. Ma, J. Duan, F. Kong, X. Shi, K. Xu, "Exposing the fake: Effective diffusion-generated images detection," *arXiv preprint arXiv:2307.06272*, 2023.

[81] J. Huertas-Tato, A. Martín, J. Fierrez, D. Camacho, "Fusing cnns and statistical indicators to improve image classification," *Information Fusion*, vol. 79, pp. 174–187, 2022.

[82] P. Lorenz, R. L. Durall, J. Keuper, "Detecting images generated by deep diffusion models using their local intrinsic dimensionality," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 448–459.

[83] L. Guarnera, O. Giudice, S. Battiato, "Level up the deepfake detection: a method to effectively discriminate images generated by gan architectures and diffusion models," *arXiv preprint arXiv:2303.00608*, 2023.

[84] D. A. Coccomini, A. Esuli, F. Falchi, C. Gennaro, G. Amato, "Detecting images generated by diffusers," *PeerJ Computer Science*, vol. 10, p. e2127, 2024.

[85] U. Ojha, Y. Li, Y. J. Lee, "Towards universal fake image detectors that generalize across generative models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 24480–24489.

[86] M. Mathys, M. Willi, R. Meier, "Synthetic photography detection: A visual guidance for identifying synthetic images created by ai," *arXiv preprint arXiv:2408.06398*, 2024.

[87] C. Tan, R. Tao, H. Liu, G. Gu, B. Wu, Y. Zhao, Y. Wei, "C2p-clip: Injecting category common prompt in clip to enhance generalization in deepfake detection," *arXiv preprint arXiv:2408.09647*, 2024.

[88] M. Keita, W. Hamidouche, H. B. Eutamene, Hadid, A. Taleb-Ahmed, "Bi-lora: A vision- language approach for synthetic image detection," *Pattern Recognition*, 2024. Preprint available at https://github.com/Mamadou-Keita/VLM-DETECT.

[89] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, D. J. Fleet, "Video diffusion models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 8633–8646, 2022.

[90] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet, *et al.*, "Imagen video: High definition video generation with diffusion models," *arXiv preprint arXiv:2210.02303*, 2022.

[91] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni, *et al.*, "Make- a-video: Text-to-video generation without text-video data," *arXiv preprint arXiv:2209.14792*, 2022.

[92] D. Zhou, W. Wang, H. Yan, W. Lv, Y. Zhu, J. Feng, "Magicvideo: Efficient video generation with latent diffusion models," *arXiv preprint arXiv:2211.11018*, 2022.

[93] L. Yu, J. Lezama, N. B. Gundavarapu, L. Versari, K. Sohn, D. Minnen, Y. Cheng, A. Gupta, X. Gu, A. G. Hauptmann, *et al.*, "Language model beats diffusion– tokenizer is key to visual generation," *arXiv preprint arXiv:2310.05737*, 2023.

[94] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, M. Tagliasacchi, "Soundstream: An end-to-end neural audio codec," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 495–507, 2021.

[95] J. Huertas-Tato, A. Martín, D. Camacho, "Understanding writing style in social media with a supervised contrastively pre-trained transformer," *Knowledge-Based Systems*, vol. 296, p. 111867, 2024.

[96] R. Girdhar, M. Singh, A. Brown, Q. Duval, S. Azadi, S. S. Rambhatla, A. Shah, X. Yin, D. Parikh, I. Misra, "Emu video: Factorizing text-to-video generation by explicit image conditioning," *arXiv preprint arXiv:2311.10709*, 2023.

[97] Y. Wang, X. Chen, X. Ma, S. Zhou, Z. Huang, Y. Wang, C. Yang, Y. He, J. Yu, P. Yang, *et al.*, "Lavie: High-quality video generation with cascaded latent diffusion models," *arXiv preprint arXiv:2309.15103*, 2023.

[98] W. Menapace, A. Siarohin, I. Skorokhodov, E. Deyneka, T.-S. Chen, A. Kag, Y. Fang, A. Stoliar, E. Ricci, J. Ren, *et al.*, "Snap video: Scaled spatiotemporal transformers for text-to-video synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7038–7048.

[99] T. Karras, M. Aittala, T. Aila, S. Laine, "Elucidating the design space of diffusion-based generative models," *Advances in neural information processing systems*, vol. 35, pp. 26565–26577, 2022.

[100] T. Chen, L. Li, "Fit: Far-reaching interleaved transformers," *arXiv preprint arXiv:2305.12689*, 2023.

[101] H. Chen, Y. Zhang, X. Cun, M. Xia, X. Wang, C. Weng, Y. Shan, "Videocrafter2: Overcoming data limitations for high-quality video diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7310–7320.

[102] X. Ma, Y. Wang, G. Jia, X. Chen, Z. Liu, Y.-F. Li, C. Chen, Y. Qiao, "Latte: Latent diffusion transformer for video generation," *arXiv preprint arXiv:2401.03048*, 2024.

[103] X. Li, W. Chu, Y. Wu, W. Yuan, F. Liu, Q. Zhang, F. Li, H. Feng, E. Ding, J. Wang, "Videogen: A reference-guided latent diffusion approach for high definition text-to-video generation," *arXiv preprint arXiv:2309.00398*, 2023.

[104] C. Wu, L. Huang, Q. Zhang, B. Li, L. Ji, F. Yang, G. Sapiro, N. Duan, "Godiva: Generating open-domain videos from natural descriptions," *arXiv preprint arXiv:2104.14806*, 2021.

[105] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, J. Sivic, "Howto100m: Learning a text-video embedding by watching hundred million narrated video clips," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 2630–2640.

[106] W. Hong, M. Ding, W. Zheng, X. Liu, J. Tang, "Cogvideo: Large-scale pretraining for text-to-video generation via transformers," *arXiv preprint arXiv:2205.15868*, 2022.

[107] C. Wu, J. Liang, L. Ji, F. Yang, Y. Fang, D. Jiang, N. Duan, "Nüwa: Visual synthesis pre-training for neural visual world creation," in *European conference on computer vision*, 2022, pp. 720–736, Springer.

[108] C. Wu, J. Liang, X. Hu, Z. Gan, J. Wang, L. Wang, Z. Liu, Y. Fang, N. Duan, "Nuwa-infinity: Autoregressive over autoregressive generation for infinite visual synthesis," *arXiv preprint arXiv:2207.09814*, 2022.

[109] W. Yan, Y. Zhang, P. Abbeel, A. Srinivas, "Videogpt: Video generation using vq-vae and transformers," *arXiv preprint arXiv:2104.10157*, 2021.

[110] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler, K. Kreis, "Align your latents: High-resolution video synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22563–22575.

[111] H. Chen, M. Xia, Y. He, Y. Zhang, X. Cun, S. Yang, J. Xing, Y. Liu, Q. Chen, X. Wang, *et al.*, "Videocrafter1: Open diffusion models for high-quality video generation," *arXiv preprint arXiv:2310.19512*, 2023.

[112] Y. He, T. Yang, Y. Zhang, Y. Shan, Q. Chen, "Latent video diffusion models for high-fidelity long video generation," *arXiv preprint arXiv:2211.13221*, 2022.

[113] J. Wang, H. Yuan, D. Chen, Y. Zhang, X. Wang, S. Zhang, "Modelscope text-to-video technical report," *arXiv preprint arXiv:2308.06571*, 2023.

[114] A. Gupta, L. Yu, K. Sohn, X. Gu, M. Hahn, L. Fei- Fei, I. Essa, L. Jiang, J. Lezama, "Photorealistic video generation with diffusion models," *arXiv preprint arXiv:2312.06662*, 2023.

[115] R. Villegas, M. Babaeizadeh, P.-J. Kindermans, H. Moraldo, H. Zhang, M. T. Saffar, S. Castro, J. Kunze, D. Erhan, "Phenaki: Variable length video generation from open domain textual descriptions," in *International Conference on Learning Representations*, 2022.

[116] Z. Xing, Q. Dai, H. Hu, Z. Wu, Y.-G. Jiang, "Simda: Simple diffusion adapter for efficient video generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7827–7839.

[117] D. J. Zhang, J. Z. Wu, J.-W. Liu, R. Zhao, L. Ran, Y. Gu, D. Gao, M. Z. Shou, "Show-1: Marrying pixel and latent diffusion models for text-to-video generation," *arXiv preprint arXiv:2309.15818*, 2023.

[118] L. Khachatryan, A. Movsisyan, V. Tadevosyan, R. Henschel, Z. Wang, S. Navasardyan, H. Shi, "Text2video-zero: Text-to-image diffusion models are zero-shot video generators," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15954–15964.

[119] W. Weng, R. Feng, Y. Wang, Q. Dai, C. Wang, D. Yin, Z. Zhao, K. Qiu, J. Bao, Y. Yuan, *et al.*, "Art-v: Auto-regressive text-to-video generation with diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7395–7405.

[120] F. Shi, J. Gu, H. Xu, S. Xu, W. Zhang, L. Wang, "Bivdiff: A training-free framework for general-purpose video synthesis via bridging image and

video diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024,pp. 7393–7402.

[121] Z. Qing, S. Zhang, J. Wang, X. Wang, Y. Wei, Y. Zhang, C. Gao, N. Sang, "Hierarchical spatio- temporal decoupling for text-to-video generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 6635–6645.

[122] R. Wu, L. Chen, T. Yang, C. Guo, C. Li, X. Zhang, "Lamp: Learn a motion pattern for few-shot-based video generation," *arXiv preprint arXiv:2310.10769*, 2023.

[123] X. Guo, M. Zheng, L. Hou, Y. Gao, Y. Deng, C. Ma, W. Hu, Z. Zha, H. Huang, P. Wan, *et al.*, "I2v-adapter: A general image-to-video adapter for video diffusion models," *arXiv preprint arXiv:2312.16693*, 2023.

[124] D. J. Zhang, D. Li, H. Le, M. Z. Shou, C. Xiong, D. Sahoo, "Moonshot: Towards controllable video generation and editing with multimodal conditions," *arXiv preprint arXiv:2401.01827*, 2024.

[125] L. Gong, Y. Zhu, W. Li, X. Kang, B. Wang, T. Ge, B. Zheng, "Atomovideo: High fidelity image-to-video generation," *arXiv preprint arXiv:2403.01800*, 2024.

[126] X. Shi, Z. Huang, F.-Y. Wang, W. Bian, D. Li, Y. Zhang, M. Zhang, K. C. Cheung, S. See, H. Qin, *et al.*, "Motion-i2v: Consistent and controllable image-to- video generation with explicit motion modeling," in *ACM SIGGRAPH 2024 Conference Papers*, 2024, pp. 1–11.

[127] W. Ren, H. Yang, G. Zhang, C. Wei, X. Du, S. Huang, W. Chen, "Consisti2v: Enhancing visual consistency for image-to-video generation," *arXiv preprint arXiv:2402.04324*, 2024.

[128] C. Shen, Y. Gan, C. Chen, X. Zhu, L. Cheng, T. Gao, J. Wang, "Decouple content and motion for conditional image-to-video generation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, 2024, pp. 4757–4765.

[129] L. Hu, "Animate anyone: Consistent and controllable image-to-video synthesis for character animationmagic," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8153–8163.

[130] Z. Xu, J. Zhang, J. H. Liew, H. Yan, J.-W. Liu, C. Zhang, J. Feng, M. Z. Shou, "Magicanimate: Temporally consistent human image animation using diffusion model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1481–1490.

[131] M. Dorkenwald, T. Milbich, A. Blattmann, R. Rombach, K. G. Derpanis, B. Ommer, "Stochastic image-to-video synthesis using cinns," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3742–3753.

[132] H. Ni, C. Shi, K. Li, S. X. Huang, M. R. Min, "Conditional image-to-video generation with latent flow diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 18444–18455.

[133] C. Wang, J. Gu, P. Hu, S. Xu, H. Xu, X. Liang, "Dreamvideo: High-fidelity image-to-video generation with image retention and text guidance," *arXiv preprint arXiv:2312.03018*, 2023.

[134] S. Zhang, J. Wang, Y. Zhang, K. Zhao, H. Yuan, Z. Qin, X. Wang, D. Zhao, J. Zhou, "I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models," *arXiv preprint arXiv:2311.04145*, 2023.

[135] A. Blattmann, T. Milbich, M. Dorkenwald, B. Ommer, "Understanding object dynamics for interactive image-to-video synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5171–5181.

[136] W. Menapace, S. Lathuiliere, S. Tulyakov, A. Siarohin, E. Ricci, "Playable video generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10061–10070.

[137] H. Wang, M. Huang, D. Wu, Y. Li, W. Zhang, "Supervised video-to-video synthesis for single human pose transfer," *IEEE Access*, vol. 9, pp. 17544–17556, 2021.

[138] L. Zhuo, G. Wang, S. Li, W. Wu, Z. Liu, "Fast- vid2vid: Spatial-temporal compression for video-to- video synthesis," in *European Conference on Computer Vision*, 2022, pp. 289–305, Springer.

[139] S. Yang, Y. Zhou, Z. Liu, C. C. Loy, "Rerender a video: Zero-shot text-guided video-to-video translation," in *SIGGRAPH Asia 2023 Conference Papers*, 2023, pp. 1–11.

[140] W. Wang, Y. Jiang, K. Xie, Z. Liu, H. Chen, Y. Cao, X. Wang, C. Shen, "Zero-shot video editing using off-the-shelf image diffusion models," *arXiv preprint arXiv:2303.17599*, 2023.

[141] C. Qi, X. Cun, Y. Zhang, C. Lei, X. Wang, Y. Shan, Q. Chen, "Fatezero: Fusing attentions for zero- shot text-based video editing," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15932–15942.

[142] E. Molad, E. Horwitz, D. Valevski, A. R. Acha, Y. Matias, Y. Pritch, Y. Leviathan, Y. Hoshen, "Dreamix: Video diffusion models are general video editors," *arXiv preprint arXiv:2302.01329*, 2023.

[143] Z. Hu, D. Xu, "Videocontrolnet: A motion-guided video-to-video translation framework by using diffusion model with controlnet," *arXiv preprint arXiv:2307.14073*, 2023.

[144] F. Liang, B. Wu, J. Wang, L. Yu, K. Li, Y. Zhao, I. Misra, J.-B. Huang, P. Zhang, P. Vajda, *et al.*, "Flowvid: Taming imperfect optical flows for consistent video- to-video synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8207–8216.

[145] B. Wu, C.-Y. Chuang, X. Wang, Y. Jia, K. Krishnakumar, T. Xiao, F. Liang, L. Yu, P. Vajda, "Fairy: Fast parallelized instruction-guided video- to-video synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8261–8270.

[146] M. Ku, C. Wei, W. Ren, H. Yang, W. Chen, "Anyv2v: A plug-and-play framework for any video-to-video editing tasks," *arXiv preprint arXiv:2403.14468*, 2024.

[147] W. Ouyang, Y. Dong, L. Yang, J. Si, X. Pan, "I2vedit: First-frame-guided video editing via image-to-video diffusion models," *arXiv preprint arXiv:2405.16537*, 2024.

[148] H. Ouyang, Q. Wang, Y. Xiao, Q. Bai, J. Zhang, K. Zheng, X. Zhou, Q. Chen, Y. Shen, "Codef: Content deformation fields for temporally consistent video processing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8089–8099.

[149] Y. Gu, Y. Zhou, B. Wu, L. Yu, J.-W. Liu, R. Zhao, J. Z. Wu, D. J. Zhang, M. Z. Shou, K. Tang, "Videoswap: Customized video subject swapping with interactive semantic point correspondence," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7621–7630.

[150] J. Bai, T. He, Y. Wang, J. Guo, H. Hu, Z. Liu, J. Bian, "Uniedit: A unified tuning-free framework for video motion and appearance editing," *arXiv preprint arXiv:2402.13185*, 2024.

[151] Y. Hu, C. Luo, Z. Chen, "Make it move: controllable image-to-video generation with text descriptions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18219–18228.

[152] Y. Guo, C. Yang, A. Rao, Z. Liang, Y. Wang, Y. Qiao, M. Agrawala, D. Lin, B. Dai, "Animatediff: Animate your personalized text-to-image diffusion models without specific tuning," *arXiv preprint arXiv:2307.04725*, 2023.

[153] S. Yin, C. Wu, H. Yang, J. Wang, X. Wang, M. Ni, Z. Yang, L. Li, S. Liu, F. Yang, *et al.*, "Nuwa-xl: Diffusion over diffusion for extremely long video generation," *arXiv preprint arXiv:2303.12346*, 2023.

[154] P. Esser, J. Chiu, P. Atighehchian, J. Granskog, Germanidis, "Structure and content-guided video synthesis with diffusion modelss," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7346–7356.

[155] S. Yin, C. Wu, J. Liang, J. Shi, H. Li, G. Ming, N. Duan, "Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory," *arXiv preprint arXiv:2308.08089*, 2023.

[156] X. Wang, H. Yuan, S. Zhang, D. Chen, J. Wang, Y. Zhang, Y. Shen, D. Zhao, J. Zhou, "Videocomposer: Compositional video synthesis with motion controllability," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[157] H. Ni, B. Egger, S. Lohit, A. Cherian, Y. Wang, T. Koike-Akino, S. X. Huang, T. K. Marks, "Ti2v- zero: Zero-shot image conditioning for text-to-video diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9015–9025.

[158] C. Nash, J. Carreira, J. Walker, I. Barr, A. Jaegle, M. Malinowski, P. Battaglia, "Transframer: Arbitrary frame prediction with generative models," *arXiv preprint arXiv:2203.09494*, 2022.

[159] D. S. Vahdati, T. D. Nguyen, A. Azizpour, M. C. Stamm, "Beyond deepfake images: Detecting ai- generated videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4397–4408.

[160] P. He, L. Zhu, J. Li, S. Wang, H. Li, "Exposing ai- generated videos: A

benchmark dataset and a local- and-global temporal defect based detection method," *arXiv preprint arXiv:2405.04133*, 2024.

[161] Z. Peng, L. Dong, H. Bao, Q. Ye, F. Wei, "Beit v2: Masked image modeling with vector-quantized visual tokenizers," *arXiv preprint arXiv:2208.06366*, 2022.

[162] H. Chen, Y. Hong, Z. Huang, Z. Xu, Z. Gu, Y. Li, J. Lan, H. Zhu, J. Zhang, W. Wang, *et al.*, "Demamba: Ai- generated video detection on million-scale genvideo benchmark," *arXiv preprint arXiv:2405.19707*, 2024.

[163] J. Bai, M. Lin, G. Cao, "Ai-generated video detection via spatio-temporal anomaly learning," *arXiv preprint arXiv:2403.16638*, 2024.

[164] L. Ma, J. Zhang, H. Deng, N. Zhang, Y. Liao, H. Yu, "Decof: Generated video detection via frame consistency," *arXiv preprint arXiv:2402.02085*, 2024.

[165] L. Ji, Y. Lin, Z. Huang, Y. Han, X. Xu, J. Wu, C. Wang, Z. Liu, "Distinguish any fake videos: Unleashing the power of large-scale data and motion features," *arXiv preprint arXiv:2405.15343*, 2024.

[166] H. Xu, J. Zhang, J. Cai, H. Rezatofighi, D. Tao, "Gmflow: Learning optical flow via global matching," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 8121–8130.

[167] Q. Liu, P. Shi, Y.-Y. Tsai, C. Mao, J. Yang, "Turns out i'm not real: Towards robust detection of ai-generated videos," *arXiv preprint arXiv:2406.09601*, 2024.

[168] J. Ricker, S. Damm, T. Holz, A. Fischer, "Towards the detection of diffusion model deepfakes," *arXiv preprint arXiv:2210.14571*, 2022.

[169] H. Song, S. Huang, Y. Dong, W.-W. Tu, "Robustness and generalizability of deepfake detection: A study with diffusion models," *arXiv preprint arXiv:2309.02218*, 2023.

[170] L. Papa, L. Faiella, L. Corvitto, L. Maiano, I. Amerini, "On the use of stable diffusion for creating realistic faces: From generation to detection," in *2023 11th International Workshop on Biometrics and Forensics (IWBF)*, 2023, pp. 1–6, IEEE.

[171] Y. Wang, Z. Huang, X. Hong, "Benchmarking deepart detection," *arXiv preprint arXiv:2302.14475*, 2023.

[172] Z. Sha, Z. Li, N. Yu, Y. Zhang, "De-fake: Detection and attribution of fake images generated by text-to-image generation models," in *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, 2023, pp. 3418–3432.

[173] Z. Xi, W. Huang, K. Wei, W. Luo, P. Zheng, "Ai- generated image detection using a cross-attention enhanced dual-stream network," in *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2023, pp. 1463– 1470, IEEE.

[174] M. A. Rahman, B. Paul, N. H. Sarker, Z. I. A. Hakim, S. A. Fattah, "Artifact: A large-scale dataset with artificial and factual images for generalizable and robust synthetic image detection," in *2023 IEEE International Conference on Image Processing (ICIP)*, 2023, pp. 2200–2204, IEEE.

[175] S. Jia, M. Huang, Z. Zhou, Y. Ju, J. Cai, S. Lyu, "Autosplice: A text-prompt manipulated image dataset for media forensics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 893–903.

[176] X. Guo, X. Liu, Z. Ren, S. Grosz, I. Masi, X. Liu, "Hierarchical fine-grained image forgery detection and localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3155–3165.

[177] G. Zingarini, D. Cozzolino, R. Corvi, G. Poggi, L. Verdoliva, "M3dsynth: A dataset of medical 3d images with ai-generated local manipulations," *arXiv preprint arXiv:2309.07973*, 2023.

[178] R. Shao, T. Wu, Z. Liu, "Detecting and grounding multi-modal media manipulation," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 6904–6913.

[179] R. Amoroso, D. Morelli, M. Cornia, L. Baraldi, A. Del Bimbo, R. Cucchiara, "Parents and children: Distinguishing multimodal deepfakes from natural images," *arXiv preprint arXiv:2304.00500*, 2023.

[180] H. Cheng, Y. Guo, T. Wang, L. Nie, M. Kankanhalli, "Diffusion facial forgery detection," *arXiv preprint arXiv:2401.15859*, 2024.

[181] J. J. Bird, A. Lotfi, "Cifake: Image classification and explainable identification of ai-generated synthetic images," *IEEE Access*, vol. 12, pp. 15642–15650, 2024, doi: 10.1109/ACCESS.2024.3356122.

[182] M. Zhu, H. Chen, Q. Yan, X. Huang, G. Lin, W. Li, Z. Tu, H. Hu, J. Hu, Y. Wang, "Genimage: A million- scale benchmark for detecting ai-generated image," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[183] Z. Lu, D. Huang, L. Bai, J. Qu, C. Wu, X. Liu, W. Ouyang, "Seeing is not always believing: benchmarking human and model perception of ai-generated images," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[184] Y. Hong, J. Zhang, "Wildfake: A large-scale challenging dataset for ai-generated images detection," *arXiv preprint arXiv:2402.11843*, 2024.

[185] S. Changpinyo, P. Sharma, N. Ding, R. Soricut, "Conceptual 12m: Pushing web-scale image-text pre- training to recognize long-tail visual concepts," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 3558–3568.

[186] P. Sharma, N. Ding, S. Goodman, R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2556–2565.

[187] K. Srinivasan, K. Raman, J. Chen, M. Bendersky, M. Najork, "Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning," in *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, 2021, pp. 2443–2449.

[188] K. Desai, G. Kaul, Z. Aysola, J. Johnson, "Redcaps: Web-curated image-text data created by the people, for the people," *arXiv preprint arXiv:2111.11431*, 2021.

[189] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, *et al.*, "Laion-5b: An open large-scale dataset for training next generation image- text models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25278–25294, 2022.

[190] Z. J. Wang, E. Montoya, D. Munechika, H. Yang, Hoover, D. H. Chau, "Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models," *arXiv preprint arXiv:2210.14896*, 2022.

[191] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, J. Xiao, "Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop," *arXiv preprint arXiv:1506.03365*, 2015.

[192] D.-T. Dang-Nguyen, C. Pasquini, V. Conotter, G. Boato, "Raise: A raw images dataset for digital image forensics," in *Proceedings of the 6th ACM multimedia systems conference*, 2015, pp. 219–224.

[193] Clip-interrogator, "Clip-interrogator," 2022. Available: https://github.com/pharmapsychotic/ clip-interrogator.

[194] ALASKA, "Alaska." https://alaska.utt.fr/. Accessed: 2024-08-04.

[195] OpenAI, "Dall·e 2." https://openai.com/product/dall-e-2. Accessed: 2024-08-04.

[196] DreamStudio, "Dreamstudio." https://beta. dreamstudio.ai/generate. Accessed: 2024-08-04.

[197] A. Krizhevsky, G. Hinton, *et al.*, "Learning multiple layers of features from tiny images." https://www.cs.utoronto.ca/~kriz/learning-features-2009-TR.pdf, 2009.

[198] R. Zellers, X. Lu, J. Hessel, Y. Yu, J. S. Park, J. Cao, A. Farhadi, Y. Choi, "Merlot: Multimodal neural script knowledge models," *Advances in neural information processing systems*, vol. 34, pp. 23634– 23651, 2021.

[199] M. Bain, A. Nagrani, G. Varol, A. Zisserman, "Frozen in time: A joint video and image encoder for end- to-end retrieval," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 1728–1738.

[200] H. Zhu, W. Wu, W. Zhu, L. Jiang, S. Tang, L. Zhang, Z. Liu, C. C. Loy, "Celebv-hq: A large-scale video facial attributes dataset," in *European conference on computer vision*, 2022, pp. 650–667, Springer.

[201] H. Xue, T. Hang, Y. Zeng, Y. Sun, B. Liu, H. Yang, J. Fu, B. Guo, "Advancing high-resolution video-language representation with large-scale video transcriptions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5036–5045.

[202] Y. Wang, Y. He, Y. Li, K. Li, J. Yu, X. Ma, X. Li, G. Chen, X. Chen, Y. Wang, *et al.*, "Internvid: A large-scale video-text dataset for multimodal understanding and generation," *arXiv preprint arXiv:2307.06942*, 2023.

[203] J. Yu, H. Zhu, L. Jiang, C. C. Loy, W. Cai, W. Wu, "Celebv-text: A large-scale facial text-video dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14805–14814.

[204] W. Wang, H. Yang, Z. Tuo, H. He, J. Zhu, J. Fu, J. Liu, "Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation," https://openreview.net/forum?id=dUDwK38MVC, 2023.

[205] H. Xu, Q. Ye, X. Wu, M. Yan, Y. Miao, J. Ye, G. Xu, A. Hu, Y. Shi, G. Xu, *et al.*, "Youku-mplug: A 10 million large-scale chinese video-language dataset for pre-training and benchmarks," *arXiv preprint arXiv:2306.04362*, 2023.

[206] W. Wang, Y. Yang, "Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models," *arXiv preprint arXiv:2403.06098*, 2024.

[207] X. Ju, Y. Gao, Z. Zhang, Z. Yuan, X. Wang, A. Zeng, Y. Xiong, Q. Xu, Y. Shan, "Miradata: A large-scale video dataset with long durations and structured captions," *arXiv preprint arXiv:2407.06358*, 2024.

[208] T.-S. Chen, A. Siarohin, W. Menapace, E. Deyneka, H.-W. Chao, B. E. Jeon, Y. Fang, H.-Y. Lee, J. Ren, M.-H. Yang, *et al.*, “Panda-70m: Captioning 70m videos with multiple cross-modality teachers,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13320–13331.

[209] S. Chen, H. Li, Q. Wang, Z. Zhao, M. Sun, X. Zhu, J. Liu, “Vast: A vision-audio-subtitle-text omni- modality foundation model and dataset,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[210] R. Girdhar, D. Ramanan, “Cater: A diagnostic dataset for compositional actions and temporal reasoning,” *arXiv preprint arXiv:1910.04744*, 2019.

[211] J. Wang, Z. Yang, X. Hu, L. Li, K. Lin, Z. Gan, Z. Liu, C. Liu, L. Wang, “Git: A generative image-to-text transformer for vision and language,” *arXiv preprint arXiv:2205.14100*, 2022.

[212] L. Han, J. Ren, H.-Y. Lee, F. Barbieri, K. Olszewski, S. Minaee, D. Metaxas, S. Tulyakov, “Show me what and tell me how: Video synthesis via multimodal conditioning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3615–3625.

[213] J. Xu, T. Mei, T. Yao, Y. Rui, “Msr-vtt: A large video description dataset for bridging video and language,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5288–5296.

[214] D. Chen, W. B. Dolan, “Collecting highly parallel data for paraphrase evaluation,” in *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, 2011, pp. 190–200.

[215] M. Monfort, S. Jin, A. Liu, D. Harwath, R. Feris, J. Glass, A. Oliva, “Spoken moments: Learning joint audio-visual representations from video descriptions,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14871–14881.

[216] B. C. Hosler, X. Zhao, O. Mayer, C. Chen, J. A. Shackleford, M. C. Stamm, “The video authentication and camera identification database: A new database for video forensics,” *IEEE access*, vol. 7, pp. 76937– 76948, 2019.

[217] Moonvalley, “Moonvalley - ai video generation,” 2024. [Online]. Available: https://moonvalley.ai/, Accessed: 2024-08-16.

[218] L. Yang, Y. Fan, N. Xu, “Video instance segmentation,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 5188–5197.

[219] L. Huang, X. Zhao, K. Huang, “Got-10k: A large high- diversity benchmark for generic object tracking in the wild,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 5, pp. 1562–1577, 2019.

[220] X. Shang, T. Ren, J. Guo, H. Zhang, T.-S. Chua, “Video visual relation detection,” in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1300–1308.

Hessen Bougueffa

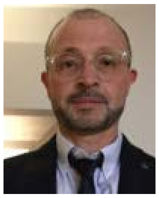
Hessen Bougueffa graduated with a Master’s degree in Telecommunication Systems in 2022. His Master’s thesis, "Deep Neural Networks Improve Radiologists’ Performance in Breast Cancer Screening," set a strong foundation in applying advanced computational techniques to real-world problems. Presently, as a Ph.D. candidate at Polytechnic Haute-de-France, Hessen is working on the development of multimodal models for content characterization in collaboration with the Martini Project. His research is carving a niche at the crossroads of machine learning and content analysis, exploring how various data types can be synergistically utilized for enhanced content understanding. Hessen’s work is expected to contribute significantly to the fields of artificial intelligence and data science, pushing the envelope in multimodal learning approaches.

Mamadou Keita

Mamadou Keita received his Engineer’s degree in Telecommunications and Computer Networks from the National Institute of Telecommunications and Information Technology in Oran, Algeria, the Master’s degree in Engineering and Innovation in Images and Networks with a specialization in Images from Sorbonne Paris Nord University, France in 2022. He is currently pursuing a Ph.D. degree in signal processing at the Institute of Electronics, Microelectronics and Nanotechnology, Polytechnic University of Hauts de France, Valenciennes, France.His research interests include image quality assessment, object detection and tracking, object segmentation, behavior analysis, medical imaging, and multimedia security.

Wassim Hamidouche

Wassim Hamidouche is a Principal Researcher at Technology Innovation Institute (TII) in Abu Dhabi, UAE. He also holds the position of Associate Professor at INSA Rennes and is a member of the Institute of Electronics and Telecommunications of Rennes (IETR), UMR CNRS 6164. He earned his Ph.D. degree in signal and image processing from the University of Poitiers, France, in 2010. From 2011 to 2012, he worked as a Research Engineer at the Canon Research Centre in Rennes, France. Additionally, he served as a researcher at the IRT b<>com research Institute in Rennes from 2017 to 2022. He has over 180 papers published in the field of image processing and computer vision. His research interests encompass various areas, including video coding, the design of software and hardware circuits and systems for video coding standards, image quality assessment, and multimedia security.



Abdelmalik Taleb-Ahmed

Abdelmalik Taleb-Ahmed received in 1992 PhD in electronics and microelectronics from université des Sciences et Technologies de Lille 1. He was associate professor in Calais until 2004. He joined the Université Polytechnique des Hauts de France in 2004, where he is presently Full Professor. He joined the laboratory IEMN DOAE. his research focused on computer vision and artificial intelligence and machine vision. His research interests include segmentation, classification, data fusion, pattern recognition, computer vision, and machine learning, with applications in biometrics, video surveillance, autonomous driving, and medical imaging. He has (co-)authored over 225 peer-reviewed papers and (co-)supervised 20 graduate students in these areas of research. His recent research revolves mainly around: Enhanced Perception and HD Mapping in intelligent Transportation, Digitalization of Road and the Signaling, E-Health and Artificial Intelligence, pattern recognition, computer vision, and information fusion, with applications in affective computing, biometrics, medical image analysis, and video analytics and surveillance.

Helena Liz-López

Helena Liz-López is an Assistant Professor in the Department of Computer Systems Engineering at the Universidad Politécnica de Madrid (UPM) and a member of the Natural Language Processing and Deep Learning (NLP&DL) research group. She holds a degree in Biology from the Universidad Autónoma de Madrid and a master’s degree in bioinformatics and computational biology from the same university. She obtained her PhD in computer sciences from the Universidad Politécnica de Madrid in 2024, receiving the distinction of “cum laude." Her research interests include Deep Learning, Machine Learning applications in ecology and medicine, and explainable AI.

Alejandro Martin

Alejandro Martin is Associate Professor at Universidad Politécnica de Madrid. His main research interests are Deep Learning, Cybersecurity, and Natural Language Processing. He has been visiting researcher at theUniversity of Kent and the University of Córdoba. Besides has participated in an important number of international conferences as a reviewer and organizer, as a reviewer and Guest Editor in international journals, and in a large number of research projects. He is the PI of different national and international projects focused on the application AI to detect and track misinformation in social networks.

David Camacho

David Camacho received the Ph.D. degree (with Honors) in Computer Xcience from Universidad Carlos III de Madrid, in 2001. He is currently a Full Professor with Computer Systems Engineering Department, Universidad Politécnica de Madrid (UPM), Madrid, Spain, and the Head of the Applied Intelligence and Data Analysis research Group, UPM. He has authored or coauthored more than 300 journals, books, and conference papers. His research interests include machine learning (clustering/deep learning), computational intelligence (evolutionary computation, swarm intelligence), social network analysis, fake news and disinformation analysis. He has participated/led more than 60 research projects (Spanish and European: H2020, DG Justice, ISFP, and Erasmus+), related to the design and application of artificial intelligence methods for data mining and optimization for problems emerging in industrial scenarios, aeronautics, aerospace engineering, cybercrime/cyber intelligence, social networks applications, or video games among others.

Abdenour Hadid

Abdenour Hadid received his Doctor of Science in Technology degree in electrical and information engineering from the University of Oulu, Finland, in 2005. Now, he is a Professor in a Chair of excellence at Sorbonne Center for Artificial Intelligence (SCAI). His research interests include computer vision, deep learning, artificial intelligence, internet of things, autonomous driving and personalized healthcare. He has authored more than 400 papers in international conferences and journals, and served as a reviewer for many international conferences and journals. His research works have been well referenced by the research community with more 25000 citations and an H-Index of 59, according to Google Scholar. Prof. Hadid was the recipient of the prestigious “Jan Koenderink Prize” for fundamental contributions in computer vision.

unir
LA UNIVERSIDAD
EN INTERNET

Rectorado
Avenida de la Paz, 137
26006 Logroño (La Rioja)
t (+34) 941 21 02 11

www.unir.net